Journal of Molecular Graphics and Modelling 29 (2010) 481–491

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Insights into the folding pathway of the Engrailed Homeodomain protein using replica exchange molecular dynamics simulations

Shruti Koulgi, Uddhavesh Sonavane, Rajendra Joshi*

Bioinformatics Team, Scientific and Engineering Computing Group (SECG), Centre for Development of Advanced Computing (C-DAC), Pune University Campus, Pune 411007, India

ARTICLE INFO

Article history:
Received 21 July 2010
Received in revised form
17 September 2010
Accepted 21 September 2010
Available online 8 October 2010

Keywords:
Protein folding
Molecular dynamics simulations
REMD
Engrailed Homeodomain
PCA

ABSTRACT

Protein folding studies were carried out by performing microsecond time scale simulations on the ultra-fast/fast folding protein Engrailed Homeodomain (EnHD) from *Drosophila melanogaster*. It is a three-helix bundle protein consisting of 54 residues (PDB ID: 1ENH). The positions of the helices are 8–20 (Helix I), 26–36 (Helix II) and 40–53 (Helix III). The second and third helices together form a Helix-Turn-Helix (HTH) motif which belongs to the family of DNA binding proteins. The molecular dynamics (MD) simulations were performed using replica exchange molecular dynamics (REMD). REMD is a method that involves simulating a protein at different temperatures and performing exchanges at regular time intervals. These exchanges were accepted or rejected based on the Metropolis criterion. REMD was performed using the AMBER FF03 force field with the generalised Born solvation model for the temperature range 286–373 K involving 30 replicas. The extended conformation of the protein was used as the starting structure. A simulation of 600 ns per replica was performed resulting in an overall simulation time of 18 μs. The protein was seen to fold close to the native state with backbone root mean square deviation (RMSD) of 3.16 Å. In this low RMSD structure, the Helix I was partially formed with a backbone RMSD of 3.37 Å while HTH motif had an RMSD of 1.81 Å. Analysis suggests that EnHD folds to its native structure via an intermediate in which the HTH motif is formed. The secondary structure development occurs first followed by tertiary packing. The results were in good agreement with the experimental findings.

1. Introduction

Understanding protein folding has been a scientifically and computationally challenging task till date. The question, "How does an amino acid sequence dictate the structure of a protein?" has been of major interest and different theories have been put forward to answer it [1–10]. Anfinsen's experiments stated that the structural information lies in the amino acid sequence itself [1]. Levinthal proposed a hypothesis stating that a protein does not pass through every possible conformer in search of the stable one but directly chooses the optimum path [10–12]. Proteins actually fold in fraction of seconds, as it follows the optimum pathway via stable intermediates. Current folding concepts believe in the ensemble view of native, unfolded and transition states as well as the landscape view, over a pathway approach [8–12]. Experimental techniques like T-jump [13], fluorescence resonance energy transfer (FRET) [14], hydrogen exchange and nuclear magnetic resonance (NMR) [13–17] are able to characterize intermediates and interesting folding and unfolding events [3,4]. The major challenges involved in the use of molecular dynamics (MD) simulations are to explore folding landscape for fast folding proteins and give an atomic level understanding of the folding process [18–25]. The recently developed algorithms (REMD and MD using coarse grained potentials) and computational power enable one to carry out long simulations [18]. The advanced methods like REMD have significantly contributed to the understanding of the folding landscape of various ultrafast folding proteins [14,15,17–19].

In this study, REMD [26] was employed to study protein folding. This technique involves simulating protein conformers at different temperatures which are known as replicas. These replicas are further exchanged at regular time intervals. The exchanges are accepted if they satisfy the Metropolis criterion [26]. REMD helps in performing large conformational searches which is one of the most important aspects of protein folding. In case of classical MD, a conformational search is carried out by random walk in energy space, whereas REMD performs random walk in temperature space followed by the same in configuration space. The exchange from low to high temperatures prevents the system from getting trapped in one of the several local minima [15]. In REMD, multiple simulations can be performed simultaneously, hence providing speed up over classical MD, to reach the experimental time scales. The major difficulty

* Corresponding author at: Bioinformatics Team, Scientific and Engineering Computing Group (SECG), Centre for Development of Advanced Computing (C-DAC), Pune University Campus, Ganeshkhind, Pune, Maharashtra 411007, India.
Tel.: +91 20 25694084; fax: +91 20 25694084.
E-mail address: rajendra@cdac.in (R. Joshi).

1093-3263/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2010.09.007

in experimental studies of protein folding lies in the capturing of intermediates, which is possible by performing REMD simulations for long duration. In this study the protein Engrailed Homeodomain (EnHD) [27], was selected as a model to understand the protein folding events. EnHD is a three helix bundle protein consisting of 54 amino acids with a melting temperature of 328 K. The protein contains a Helix Turn Helix (HTH) motif, the signature domain of DNA binding proteins, known to fold fast and independently. Experimental methods suggest that the formation of HTH (Helices II and III) motif in EnHD corresponds to a fast folding phase comprising of a microsecond (μs), while the docking of Helix I requires about 10 μs [14]. Different models of protein folding mechanisms have been proposed, viz. nucleation–condensation, framework and diffusion–collision [13–16,28,29]. Proteins belonging to the family of Homeodomain show complete transition from concurrent (nucleation–condensation) to sequential (framework mechanism) secondary and tertiary structure formation. EnHD seems to fold by a classical diffusion–collision mechanism wherein the secondary structural elements form independently and then dock to form the tertiary structure [28]. Experimental studies reported on EnHD show that the transition state in the folding pathway is more native like and stable [16]. The major MD simulations related to EnHD have been performed at high and moderately high temperatures and are targeted to study the unfolding and its effect on the secondary structure [15,30]. These unfolding simulations revealed the reversible nature of EnHD, as back tracing the unfolding pathway led to the folded conformation of the protein [31,32]. The microscopic reversibility of EnHD has been studied by performing large scale simulations at temperatures higher than the melting temperature for EnHD. The results obtained proved that the unfolding and folding transition states are remarkably similar to each other with their paths completely deviated [31,32].

The present simulation study seeks to reach the experimental folding timescale of EnHD and characterize the transient intermediates obtained in the path of folding. The results provide an insight into the folding landscape of EnHD. The results obtained were compared with the previously reported experimental data on EnHD [13,14,27].

2. Methodology

The simulations were performed using the AMBER 10 suite of programs [33]. The initial structure of EnHD was built in *xleap* with values of phi–psi angles being 180°. The all-atom point charge FF03 force field was used to assign the parameters and the total size of the system resulted in 944 atoms. The Hawkins, Cramer and Truhlar pairwise generalised Born (GB) model was used in order to mimic the solvation effect. The surface area term was not included. The Langevin thermostat was used.

The extended polypeptide was subjected to a short minimization of 5000 steps comprising of initial 2500 steps of steepest descent and remaining 2500 steps of conjugate gradient. This minimized structure was further heated till 300 K and then equilibrated for 1 ns at the same temperature. The equilibrated structure was then subjected to REMD, with selected temperature range of 286–373 K, resulting in 30 replicas. The replicas were selected by considering an equal distribution around the melting temperature (328 K) [13,14] with difference between each replica being 3 K. Each replica was simulated for 600 ns, resulting in a total simulation time of 18 μs. Exchanges were performed every 1 ns for the initial 10 μs and then the exchange interval was reduced to 500 ps for the remaining 8 μs. The number of exchanges was increased in the latter part of the simulation to explore more conformational space. These simulations were performed on parallel computing clusters namely BIOGENE and PARAM Yuva using a maximum of 480 processors. The analyses were performed using the *ptraj* [33] and *mm-pbsa* [33] modules of AMBER. *Ptraj* was used for the root mean square deviation (RMSD) calculations and for performing the principal component analysis (PCA) [34]. The *mm-pbsa.pl* program from the *MM-PBSA* module was used to perform the free energy calculations and *NACCESS* [35] was used to calculate the solvent accessible surface area (SASA). Secondary structure analyses were performed using the *STRIDE* package [36] and *VMD* [37] was used for visualizing the trajectories. *Plotmtv* and *XMGRACE* were used for plotting the data.

3. Results and discussion

REMD simulations of EnHD resulted in generation of 30 trajectories, one for each replica. All the replicas consist of conformers which had visited different temperatures as a result of exchanges. As discussed earlier, the exchanges were scrutinized by the Metropolis criterion, obeying which the exchange was accepted. Based on this, the acceptance ratio (ratio of number of times an exchange was accepted to total no. of exchanges occurred) for each replica was obtained. The acceptance ratio for all the replicas remained constant at around 0.8. The high acceptance ratio states that the selected replicas were exchanged for a large number of times. Hence, it can be inferred that the REMD simulations were significant and they explored large conformational space.

The snapshots at every 100 ps of the trajectories were used for analysis. These analyses include RMSD, SASA, PCA and free energy calculations. In order to check the convergence of the replicas, the backbone RMSD for one of the replicas was analyzed. RMSD was calculated with reference to the native EnHD (PDB ID: 1ENH) and only backbone atoms were considered in this entire study. Fig. 1A describes the backbone RMSD for the entire structure and individual helices against time for one of the replicas. The entire protein RMSD (shown in blue in Fig. 1A) stabilized around 3–4 Å after 100 ns. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.) The Helix I RMSD (shown in red in Fig. 1A) was higher as compared to that of Helix II (shown in green) and III (shown in purple). It was observed that after 100 ns the Helix I RMSD stabilized around 3–4 Å whereas the same for Helices II and III had stabilized below 2 Å. The secondary structure growth for the same was calculated using STRIDE secondary structure prediction tool (Fig. 1B). The magenta colour corresponds to the alpha helical content. A gradual growth in the secondary structural content of the protein was observed. The temperatures which were visited by this replica ranged around 300 K throughout the simulation (Supplementary data S1(A)). In order to attain deeper insight into the convergence of these REMD simulations, implicit classical molecular dynamics (classical MD) simulation was performed at 300 K for 1 μs. The parameters such as fractional native contact and radius of gyration were plotted against time. In Fig. 1C the native contacts have been defined as the contacts where the distance between any two C^{α} atoms is less than 6.5 Å. The radius of gyration stabilized around 10.5 Å (Supplementary data S1(B)) and the fractional native contacts stabilized around 0.78 in REMD as well as the classical MD simulations. On calculating the RMSD and fractional native contacts for the REMD replicas and comparing these values with that of the classical MD simulations, it was seen that most of them showed similar results to that of the replica shown in Fig. 1. This helps to prove that most of the replicas had converged and the simulations performed in this study were significant.

The minimum RMSD values considering the entire protein obtained for the population at every temperature have been shown in Fig. 2A. These values ranged between 3 and 4 Å. The average value for the same was 3.43 Å (±0.16). Fig. 2B shows the minimum RMSD

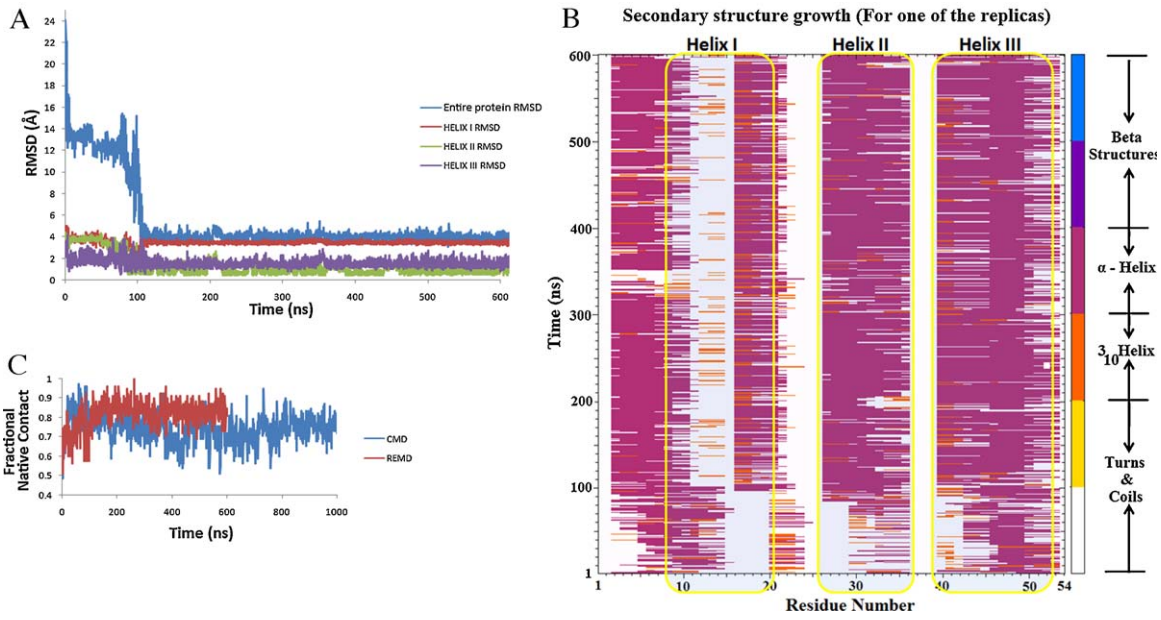

Fig. 1. Checking the convergence for REMD simulations performed. (A) Backbone RMSD for one of the replicas. (B) Secondary structure growth for the same. (C) Comparing fractional native contacts for REMD and classical MD simulations.

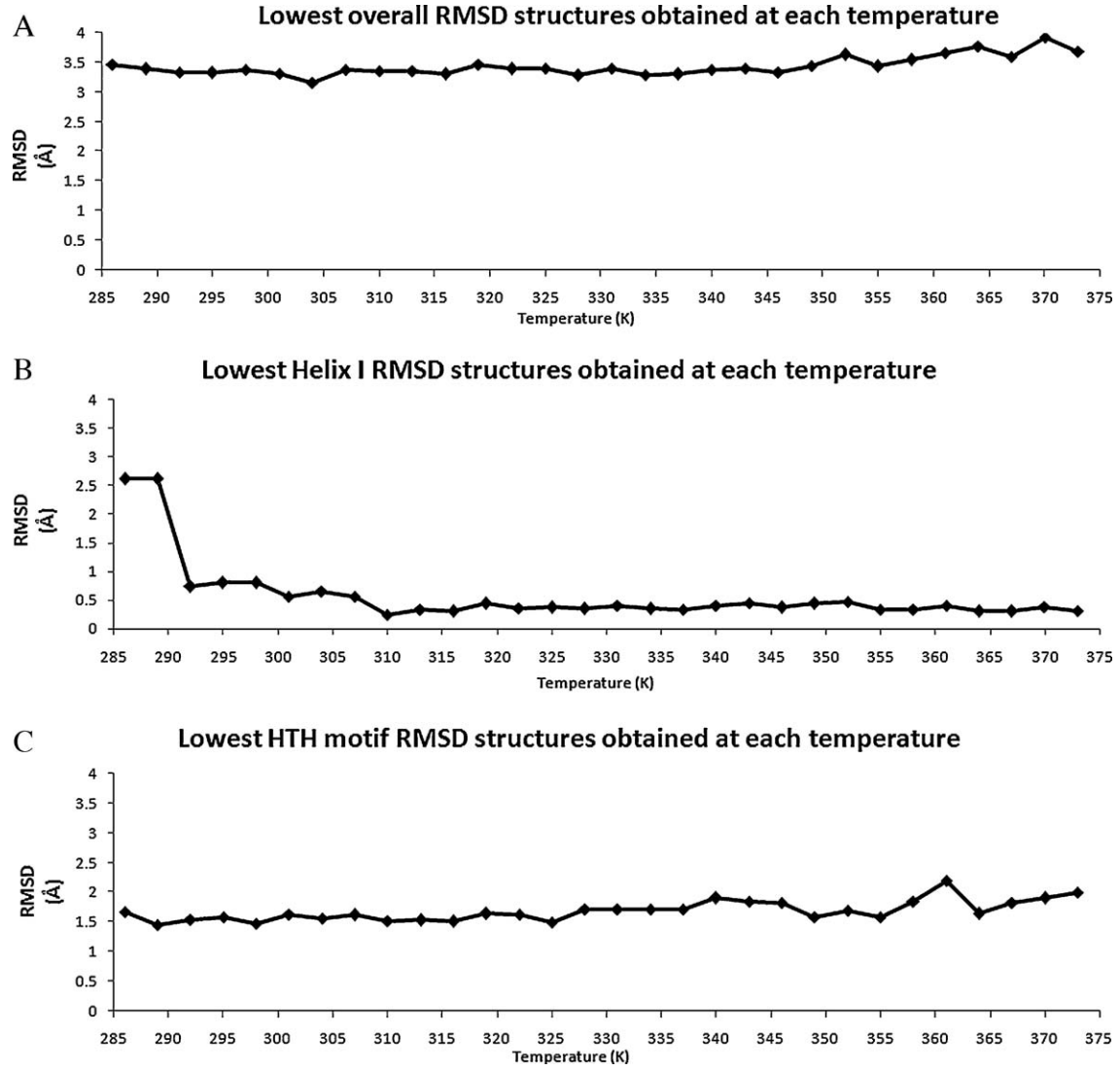

Fig. 2. Lowest RMSD at each temperature. (A) Entire protein (1–54 residues). (B) Helix I (8–20 residues). (C) Helix Turn Helix motif (26–53 residues).

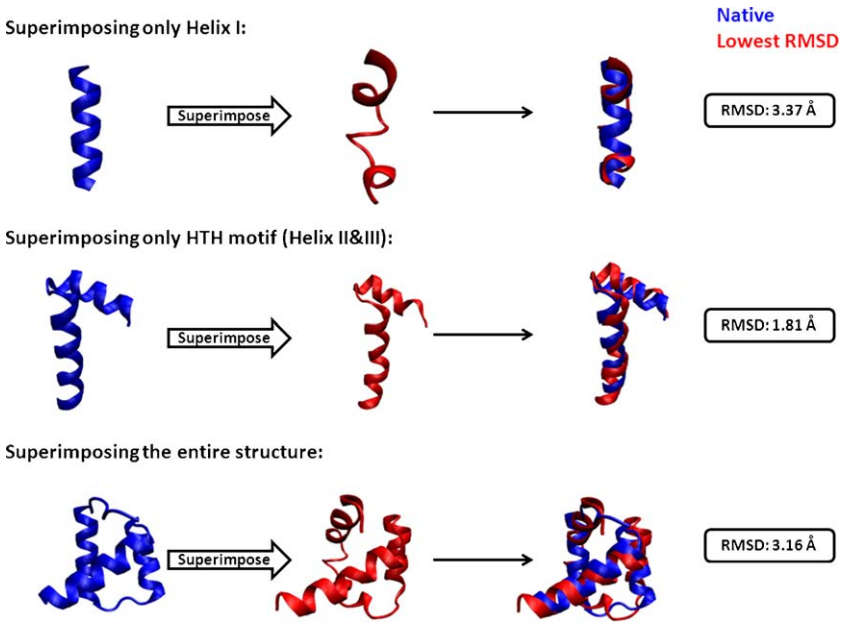

Fig. 3. Superimposed native and lowest RMSD structure.

values considering only the Helix I (8–20 residues) obtained for the population at every temperature. These values ranged below 1 Å for most of the temperatures. The average value for the same was 0.43 Å (±0.15) excluding the first two temperatures. Fig. 2C shows the minimum RMSD values considering only the HTH motif (26–53 residues) at every temperature. These values ranged below 2 Å for most of the temperatures. The average value for the same was 1.68 Å (±0.17). These plots show that at every temperature low RMSD structures are obtained and the individual helices, when considered separately, have even lower RMSD values. This shows that at every temperature near-native conformations have been obtained. On analyzing all 30 replicas, the lowest RMSD value of 3.16 Å was obtained at 297 ns where the corresponding temperature was 304 K. The Helix I RMSD for this structure was 3.37 Å and the HTH motif RMSD was 1.81 Å (Fig. 3).

3.1. Clustering based on segment-wise RMSD

EnHD was divided into two segments, A and B, segment A consisted of Helix I, and segment B consisted of the HTH motif. The structures from all the 30 trajectories were clustered by setting RMSD criterion for these two segments. Four distinct categories were formed based on the RMSD criterion defined in Table 1. The entire REMD population was segregated based on this category and it was observed that 2% *Folded*, 24% *Intermediate I*, 1% *Intermediate II* and 73% *Unfolded* population was formed. The structures from individual temperatures were analyzed to obtain the population of the different conformations adopted by the protein at these temperatures (Fig. 4A). Fig. 4A shows that the *Folded* and *Intermediate II* population were very low as compared to that of *Intermediate I* at all the temperatures. However, the *Folded* and *Intermediate II* population decreased with increase in temperature. The *Intermediate I* population was high (compared to that of *Folded* and *Intermediate II*) at each temperature and it increased with increase in temperature. It was observed that *Unfolded* population decreased with increase in temperature (data not shown). This decrease in the *Unfolded* population can be attributed to the increase in the *Intermediate I* population. A detailed study of this behavior was performed by modifying the definition of segment A, referred to as segment A′ which consisted of Helices I and II together (8–36 residues). The redefinition of segment A has been done to understand the independent folding nature of Helix I. So the segment A′ has been defined with Helix I and Helix II wherein Helix II is the overlapping region which was absent in case of segment A. Based on segment A′, a new conformation referred to as *Intermediate I′* was defined in which the RMSD of segment A′ <3.5 Å and that of segment B >3.5 Å. Fig. 4B shows that the *Intermediate I′* population increases with increase in temperature. The *Unfolded* population remained constant (data not shown) at all the temperatures. These results justify that increase in the *Intermediate I* population was responsible for the reduction in the *Unfolded* ensemble with increase in temperature. The formation of the *Intermediate II* and *Folded* conformations goes hand in hand. Hence, it can be inferred that the HTH motif (folded portion in *Intermediate II*) has an important role to play in the folding of the entire protein. The analysis for the structures obtained at temperatures 304 K (below melting temperature of EnHD) and 370 K (above melting temperature) are discussed later in this study.

Table 1
The criterion that was set for four different categories of *Folded*, *Intermediate I*, *Intermediate II* and *Unfolded*. Segment A: Helix I (8–20 residues) and segment B: HTH motif (26–53 residues).

Category	RMSD	
	Segment A	Segment B
Folded	<3.5 Å	<3.5 Å
Intermediate I	<3.5 Å	>3.5 Å
Intermediate II	>3.5 Å	<3.5 Å
Unfolded	>3.5 Å	>3.5 Å

Intermediates were formed using segment A (Helix I) and segment B (Helices II and III).

3.2. Population landscape built using segment-wise RMSD

The protein when represented in the form of segments as described earlier makes it easier to understand the important intermediates which the protein would visit before attaining its native conformation. In order to get more information about the folding pathway, a population landscape based on the segment-wise RMSD was built. Fig. 5A shows the population landscape built using the

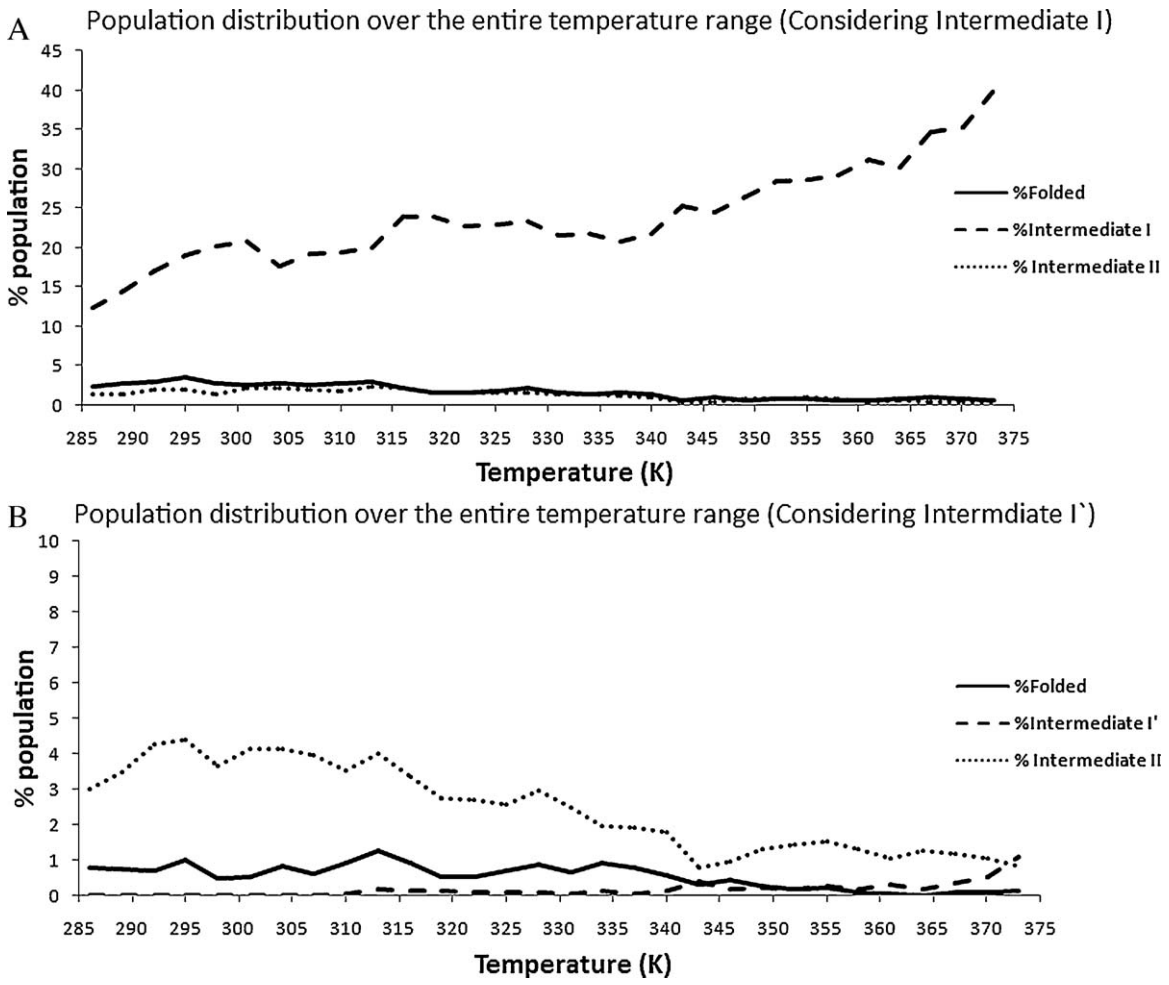

Fig. 4. Population distribution of all the structures obtained. (A) At every temperature. *Folded* (—), *Intermediate I* (– – –) and *Intermediate II* (···). (B) Same as (A) except *Intermediate I′* (– – –).

segment-wise RMSD for the structures obtained at temperature 304 K. In this plot, the RMSD of segment B has been plotted against the RMSD of segment A with the colour gradation indicating the number of structures (population). The plot has been divided into a grid in order to obtain a clear demarcation between the four distinct conformations. Regions I1, U, I2 and F correspond to *Intermediate I*, *Unfolded*, *Intermediate II* and *Folded* respectively. The plot clearly indicates a significant population of *Intermediate II* in which the RMSD of segment B is below 2.5 Å. Hence, it can be inferred that the HTH motif (folded portion of *Intermediate II*) has reached near to its native conformation. The population density for the *Unfolded* conformation was the largest followed by *Intermediate I*, as the population was widely distributed in I1 region. The population distribution was very low for *Intermediate II* and *Folded* conformations. The F region in Fig. 5A depicts the *Folded* ensemble. Segment A and segment B both showed a considerable drop in RMSD but these events did not occur simultaneously, due to which distinct population for both the intermediates were obtained. This observation led to the inference that the Helix I (folded portion of *Intermediate I*) and the HTH motif (folded portion of *Intermediate II*) are two independently folding regions of the protein. In order to study the behavior of these conformations at moderately high temperature the population landscape at 370 K was built (Supplementary data S2(A)). It was observed that the structures were scattered over the entire space with no densely populated regions. The *Unfolded* population dominated at 370 K followed by *Intermediate I*, with only a few structures lay in the region F (*Folded*) and region I2 (*Intermediate II*).

In order to study the distribution of the entire REMD population, a population landscape considering the structures obtained from all the trajectories was built (Fig. 5B). The four distinct regions (I1, U, I2 and F) could be observed and the transitions between them were also very prominent. Fig. 5B clearly indicates that the native-like *Folded* ensemble can be formed by the path passing through the *Intermediate II*. There are very few structures that enter the *Folded* region from the *Intermediate I* conformation. This information supports the earlier findings that formation of *Intermediate I* is independent and is yet to reach the native ensemble. Hence, with the help of these population landscapes it can be inferred that the formation of Helix I is independent to that of the HTH motif. The HTH motif formation further leads to the folding of the remaining protein to reach the native state. The population landscape was built for the entire REMD population considering segment A′ which has been defined earlier (Supplementary data S2(B)). Segment A was redefined as segment A′ in order to understand the independent folding nature of Helix I. It was observed that *Intermediate I′* population is very low which indicates that segment A′ had not attained the native-like conformation. On comparing the landscapes built using segment A and segment A′ it can be seen that the *Intermediate I′* population has shifted to the *Unfolded* region. Hence, the segment A′ contributes very less to the formation of an intermediate as compared to that of segment A. Helices I and II which belong to the folded portion on *Intermediate I′* do not contribute to the folding pathway of EnHD. Hence, the further analyses were performed considering segment A.

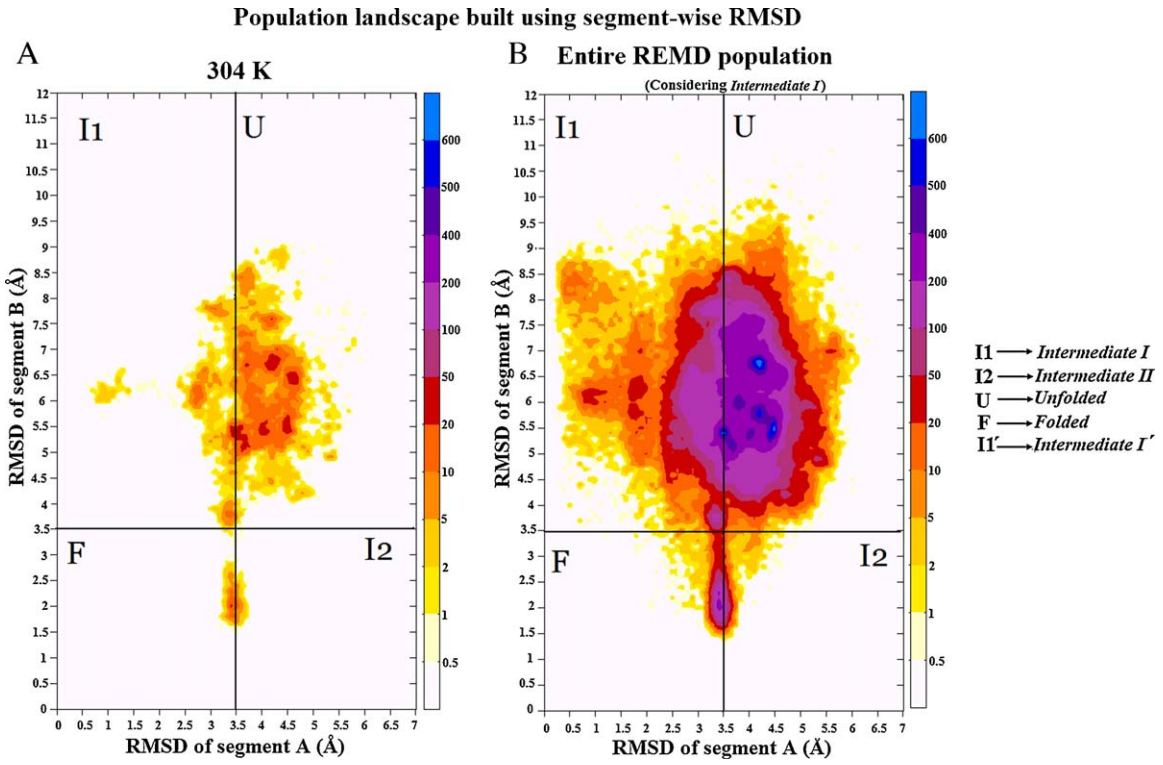

Fig. 5. Population landscape built using segment-wise RMSD. (A) For structures obtained at 304 K. (B) For structures obtained from all the trajectories (considering segment A (*Intermediate I*)).

3.3. Population landscape using principal component analysis (PCA)

PCA is a mathematical technique which employs a co-variance matrix to reduce a multidimensional data to a lower dimension. Hence, this technique can be implemented to identify the diffusive properties of a protein. The structures obtained from all the trajectories were subjected to PCA where the backbone atoms were considered as the reaction coordinates. The population landscape was built by plotting the Principal Component 1 (PC1) against the Principal Component 2 (PC2). Obtaining a well-populated cluster in any region of the PCA indicates that the projections at that point correspond to dominant local motions of the atoms in the protein. Fig. 6A shows the projection of PC1 on PC2 for structures obtained at 304 K. The colour gradation refers to the number of structures (population). These structures formed four distinct densely populated clusters and the *Folded*, *Intermediate I*, *Intermediate II* and *Unfolded* populations for each of these were calculated. Whenever for a particular pair of PC1 and PC2 values, the number of structures exceeded 200, it was considered as a cluster. Cluster-1 and Cluster-3 mainly consisted on the *Unfolded* population (99%) with the lowest RMSD value of 9.43 Å and 8.69 Å respectively. Cluster-2 showed the presence of the *Folded* (57%), *Intermediate I* (4%) and *Intermediate II* (39%) conformations where the *Intermediate I* population was less as compared to that of *Intermediate II*. On calculating the RMSD of the structures in cluster-2, the lowest RMSD of 3.24 Å was obtained. In this cluster, the *Folded* population was the largest followed by the *Intermediate II* population, which infers that the local motions facilitating the formation of *Intermediate II* also contribute to formation of native like conformation. Cluster-4 showed a small population of *Intermediate I* (24%) which states that the Helix I (folded portion of *Intermediate I*) formation has occurred in the structures of this cluster. The RMSD calculations performed on the structures of this cluster revealed that the lowest RMSD value was 10.6 Å. PCA was performed for structures obtained from all the trajectories and it was observed that the clusters formed were less populated with increase in temperature. Fig. 6B shows the projection of PC1 on PC2 for conformers obtained at moderately high temperature of 370 K and the colour gradation refers to the number of structures (population). Here, it can be clearly observed that the population distribution is scattered and there are no densely populated clusters. Only a single sparsely populated cluster was obtained. In this cluster, the *Intermediate I* was formed but the population was low as compared to that of the *Unfolded* population (90%) and a lowest backbone RMSD of 9.88 Å was obtained. This shows that local motions in EnHD get affected with increase in temperature resulting in different intermediates. The local motions occurring at lower temperatures facilitate the formation of the HTH motif (folded part of *Intermediate II*) and do not contribute much to the formation of Helix I (folded part of *Intermediate I*).

3.4. Population landscape using MM-GBSA free energies and entire protein RMSD

Free energy was calculated for structures obtained from all the temperatures using the *mm-pbsa.pl* program from the *MM-PBSA* module of AMBER 10. Fig. 7A describes the free energy distribution over the entire RMSD range at 304 K, which was plotted against the entire protein backbone RMSD. The structures were binned down to four clusters by defining a RMSD range. The range was defined as 3–4 Å, 7–8 Å, 9–10 Å, and 10–12 Å for cluster-1, cluster-2, cluster-3 and cluster-4 respectively. The average RMSD was 3.8 Å, 7.3 Å, 9.3 Å and 10.8 Å for these four clusters respectively. These clusters were segregated based on the conformations, *Folded*, *Intermediate I*, *Intermediate II* and *Unfolded* conformation. The *Folded* (55%) and *Intermediate II* (45%) population were found only in cluster-1, where the *Folded* population was larger than the *Intermediate II* population. In case of the other three clusters, *Intermediate I* and *Unfolded* populations were found, wherein the *Unfolded* population was more than *Intermediate I*. The average free energy and

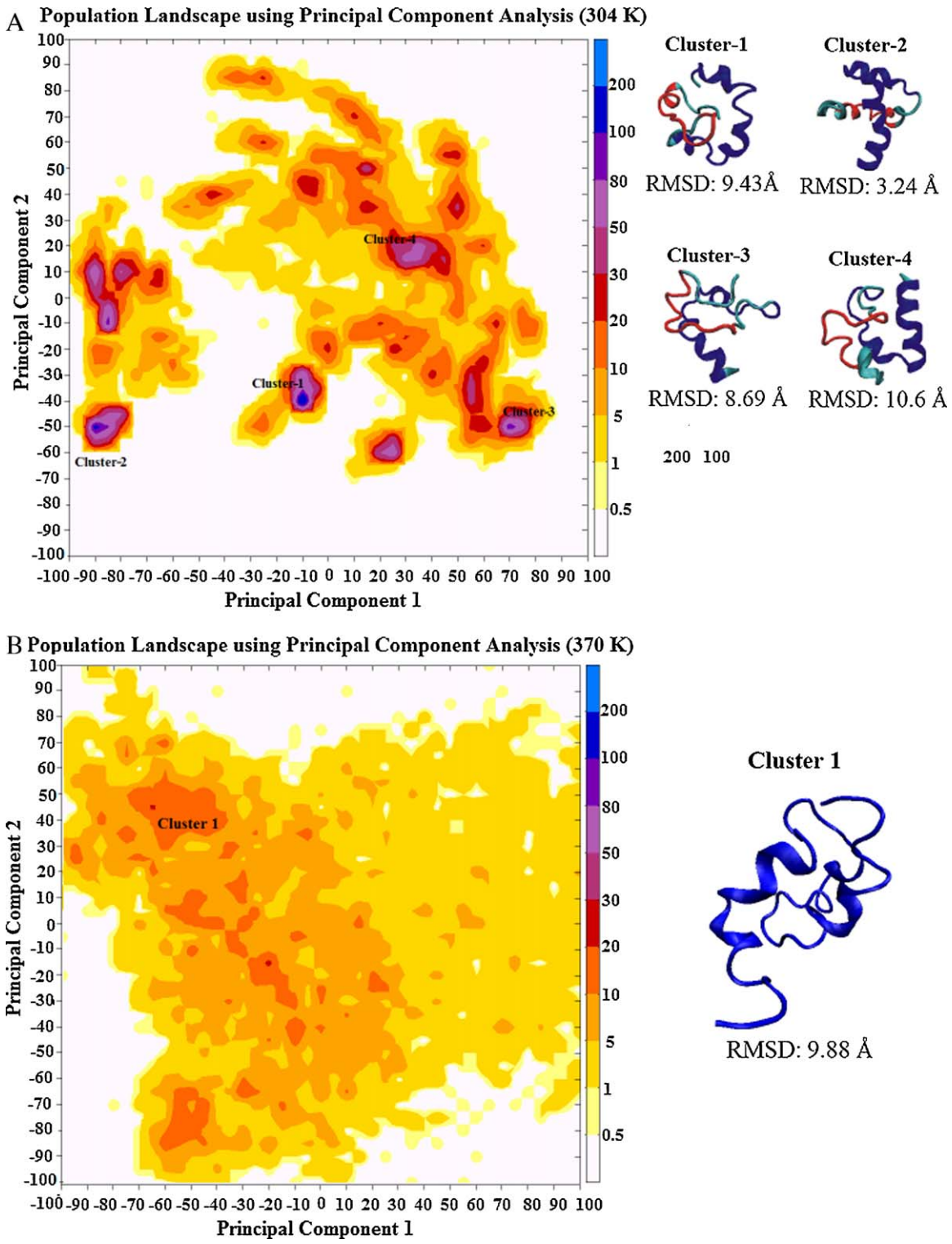

Fig. 6. Population landscape built using principal component analysis (PCA). (A) For structures obtained at 304 K. (B) For structures obtained at 370 K.

RMSD were calculated for each of these clusters as they were not spread over a large free-energy range. The plot of average free energy against average RMSD for these four clusters was calculated. The cluster containing the *Folded* conformation had the lowest free energy. The difference between the free energies for these clusters was a few kilocalories but the lowest free energy was observed for the cluster containing the native-like conformation (cluster-1). On observing the free energy distribution for conformers at higher temperatures, the cluster in the RMSD range of 3–4 Å was found to disappear with increase in temperature and after 340 K this cluster was not seen. There were only 2–5 structures which entered the *Folded* conformation with majority of them lying in high RMSD clusters. Fig. 7B describes the population distribution at a moderately high temperature of 370 K. This figure clearly indicates that only a few structures were lying in the RMSD range of 3–4 Å and densely populated clusters were formed at a RMSD of

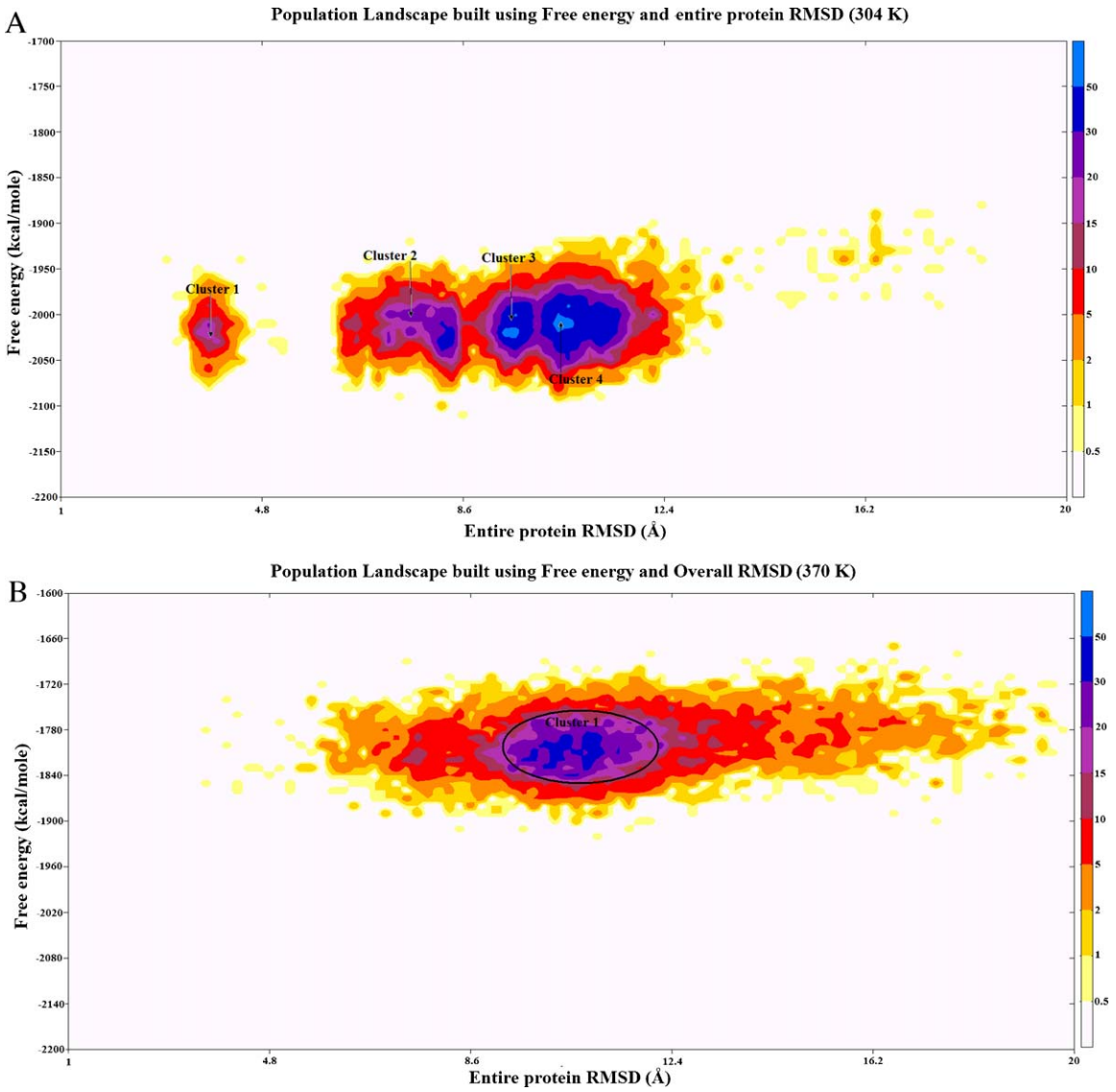

Fig. 7. Population landscape using MM-GBSA free energies and entire protein RMSD. (A) For structures obtained at 304 K. (B) For structures obtained at 370 K.

more than 10 Å. The conformers lying between 10 and 12 Å RMSD (circled region in Fig. 7B) when analyzed further, revealed that the *Intermediate I* (29%) and *Unfolded* (69%) populations were prominent. The population of *Intermediate II* (2%) though obtained was very low as compared to that of *Intermediate I* and *Unfolded* populations. The average free energy of this cluster was found to be ranging around −1800 kcal/mol which was high as compared to that of cluster-1 at 304 K (Fig. 7A) which showed free energy value around −2027 kcal/mol. These findings can be correlated to infer that *Folded* and *Intermediate II* conformations were formed at lower free energy as compared to *Intermediate I*. *Intermediate I* was formed at every temperature and was found even in regions where the free energy was high. The simultaneous occurrence of the *Folded* and *Intermediate II* population in the low RMSD and free energy clusters indicate that *Intermediate II* may have an important role to play in the formation of near to native-like conformations.

3.5. Comparison with experimental data

3.5.1. Solvent accessible surface area (SASA)

SASA represents the area of the protein exposed to the solvent and hence, indicates the packing of hydrophobic core residues. This property of a protein is very useful in understanding hydrophobic collapse which is an important event in the course of protein folding. The structure obtained with an entire protein RMSD of 3.16 Å (Fig. 3: Lowest RMSD structure obtained) was analyzed to verify the formation of the hydrophobic core. According to the X-ray studies, the structure of EnHD consists of 10 important residues which are a part of the hydrophobic core [27]. These residues are found to be buried in the native structure, making them least accessible to the solvent. The SASA values for these 10 residues in both the native (Fig. 8: shown in blue) and the lowest RMSD (Fig. 8: shown in red) structure were calculated. The differences in SASA values between the lowest RMSD structure and the native structure for each of the 10 residues were plotted against the corresponding hydrophobic residue (Fig. 8A). The plot also shows the superimposed side chains of the corresponding residues with blue indicating the native and red indicating the lowest RMSD structure. Lower variation indicates that the corresponding residues show similar hydrophobic packing as seen in the native structure. The SASA values of the last six residues, viz. LEU32, LEU36, LEU38, ILE43, TRP46 and PHE47 display significant similarity to the native structure as compared to the first four residues. These last six residues are the part of the HTH motif and it was observed that these residues were closer to

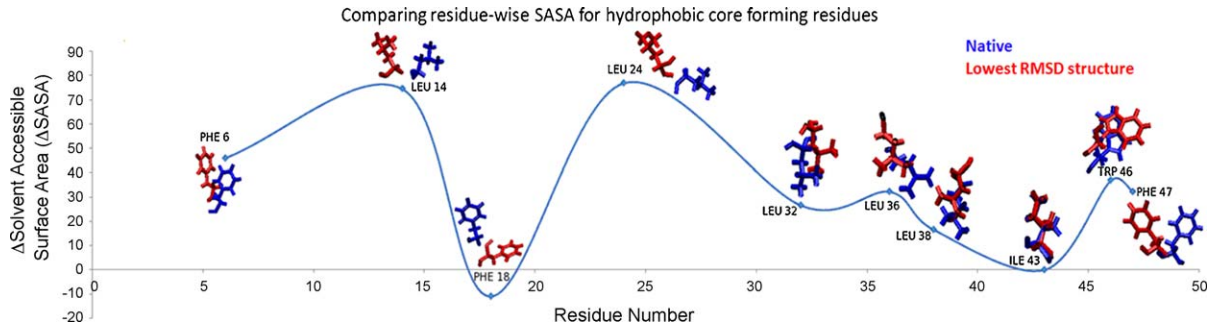

Fig. 8. Experimental comparison: hydrophobic core. The SASA difference between lowest RMSD structure and native structure for the 10 hydrophobic residues (6, 14, 18, 24, 32, 36, 38, 43, 46 and 47). Native (blue), lowest RMSD (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the native-like conformation. PHE18 which belongs to the Helix I was observed to be more buried as compared to the native. It was observed in case of the lowest RMSD structure that the Helix I was incompletely formed (Fig. 3). This can be attributed to the fact that PHE18 lies in the region corresponding to the missing helical content and shows unfavorable hydrophobic interactions.

3.5.2. Distances between critical residues

EnHD being a fast folding protein has been studied extensively in the field of protein folding. Experimental studies such as T-jump kinetics and FRET analysis have been performed in earlier studies [13,14]. T-jump experiments suggest that the HTH motif of EnHD is stable and contributes to the folding of the entire protein to its native conformation. FRET [14] analysis has been reported as a support to the T-jump kinetic experiments. According to these reported studies the distances between residues PHE6, SER8, TYR23 and ASN39 are considered as critical for the protein to fold. The distances between these residues are important as they give some information related to the positioning of the helices and folding of the same [14]. Table 2 shows the distances between these residues in the native protein as well as the lowest RMSD (Fig. 3) structure obtained. Fig. 9 describes the four distances stated in Table 2 for the native (shown in blue) and the lowest RMSD structure (shown in red). (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.) Fig. 9A describes the distance between PHE6 and ASN39. The region covered in between these residues comprised of Helices I and II. The difference in this distance (Distance 1) indicates that the Helices I and II have been partially oriented as present in the native form. Similar results are seen in case of the Distance 2 which is between SER8 and ASN39 (Fig. 9B). Fig. 9C depicts the Distance 3 which is between SER8 and TYR23 (this region comprises of Helix I). The difference in this distance was large enough which again supports the earlier finding that the Helix I was partially formed. Fig. 9D depicts the distance between the residues TYR23 and ASN39 which covers the region containing Helix II (part of the HTH motif). The difference in this distance between the native and the lowest RMSD structure was very low which indicates that the lowest RMSD structure reached a near-native conformation, considering only the HTH motif. These comparisons with the experimental data help to give a clear picture of the folding process of EnHD and hereby support the simulation data reported here.

3.6. Three-state folding pathway

EnHD, as the name suggests belongs to the Homeodomain family and experimental and simulation studies depict that these proteins show mixed behavior during their folding process. Some of the proteins are reported to follow the nucleation–condensation model and some the framework mechanism. The diffusion–collision mechanism which is also observed in this model, shows the formation of secondary structures which are stable followed by packing which involves tertiary interactions [16,28]. In this study, formation of the secondary structural elements occurred independently and the tertiary interactions were partially formed. It can be inferred that it follows the diffusion–collision mechanism as the secondary structures formed were stable and tertiary packing of the structure was partially done. Experimental results have shed more light into the fast folding phase of the HTH motif formation followed by the slow phase of Helix I docking and native structure formation [13,14]. Unfolding simulations performed on EnHD at 100 °C and 225 °C have shown disruption of helical packing [15], in particular Helix III moved away from the core and Helix II partially separated from Helix I. This suggests that unfolding pathway of EnHD tries to disrupt the HTH motif first as compared to Helix I. Hence, at moderately high temperatures the HTH motif becomes more unstable as compared to Helix I. The results reported here match with these findings as reduced population of *Intermediate II* was observed at moderately high temperature and this intermediate showed the formation of the HTH motif. In the earlier reports on simulation of the crystal structure of EnHD (PDB ID: 1ENH) performed at room temperature for 15 ns [28], it was seen that the resultant structure showed C^{α}-RMSD of 3.13 Å. This indicates that the 'folded basin' comprising of the native ensemble is large enough as the deviation is ~3 Å. Hence, it can be inferred that the lowest RMSD structure of 3.16 Å obtained at 304 K in the simulations reported may belong to the native ensemble.

In this study, the REMD simulation of 18 μs could explore the biphasic folding nature of EnHD protein. The folding landscapes built on the basis of the geometric parameters (RMSD and PCA)

Table 2
The distances between critical residues in the native protein (PDB ID: 1ENH) and lowest RMSD structure obtained.

Distance	Start residue	End residue	In native (Å)	In lowest RMSD (Å)
Distance 1	6	39	10.75	13.72
Distance 2	8	39	17.54	12.71
Distance 3	8	23	23.21	20.19
Distance 4	23	39	18.86	16.68

The distances between the critical residues have been taken from Huang et al. Fluorescence resonance energy transfer (FRET) analysis [14].

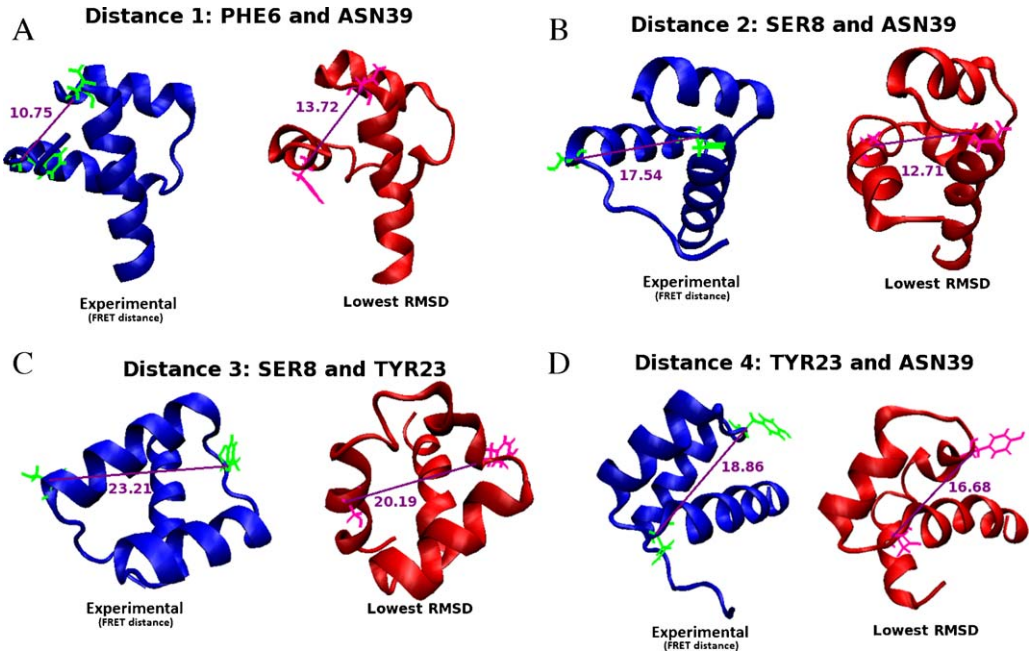

Fig. 9. Experimental comparison [14]: critical residues (measured in Å). (A) Distance between PHE6 and ASN39. (B) Distance between SER8 and ASN39. (C) Distance between SER8 and TYR23. (D) Distance between TYR23 and ASN39.

and the free energy calculations, suggests that the *Intermediate II* and *Folded* conformations coexist in majority of the clusters. These results support the fact that EnHD follows a 'three-state' folding process which involves crossing two major energy barriers to reach the native ensemble. The first energy barrier is present in the formation of the HTH motif, i.e. *Intermediate II*. This energy barrier was easily surpassed during the folding process leading to the formation of a stable HTH motif in the temperature range 286–373 K. The second energy barrier occurs during the formation and docking of Helix I which was observed in most of the replicas. In the current study, it was observed that both these barriers were surpassed in a partial manner to reach the native structure. The population landscapes show that the formation of HTH motif direct the protein further to reach the native ensemble by crossing the second energy barrier with the partial formation of Helix I and its docking to the HTH motif. Fig. 10 indicates that the path through *Intermediate I* acts like an "off pathway" in the folding process. At moderately high temperatures, in spite of the increase in the population of this intermediate, no significant increase was observed in the *Folded* population. The analysis of *Intermediate II* formation revealed that it acts like an intermediate of the "on pathway" as increase in the population of *Intermediate II* leads to simultaneous increase in the formation of the *Folded* structures. The experimental studies on EnHD [13,14,27] also support these findings and show the importance of the HTH motif in the folding process.

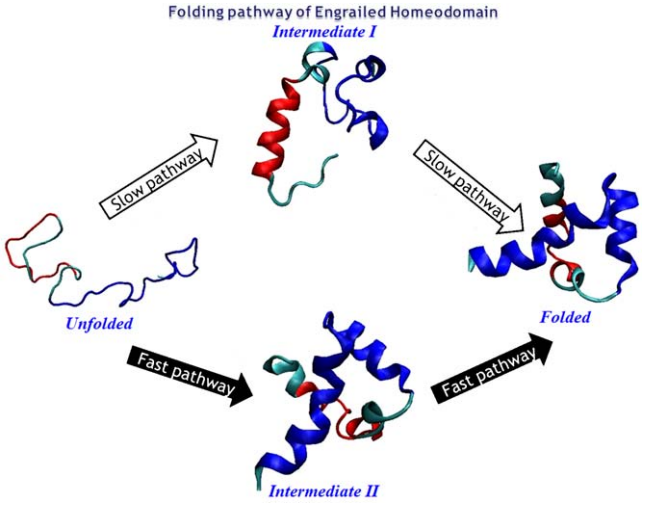

Fig. 10. The three state folding pathway.

4. Conclusion

The folding REMD simulations of Engrailed Homeodomain for 18 μs was carried out to understand its free energy landscape and to characterize the intermediates. The lowest backbone RMSD structure obtained was 3.16 Å for the entire protein. The RMSD of Helix I was 3.36 Å and that of HTH motif was 1.81 Å. The population landscape was built and divided into four different conformations, viz. *Folded*, *Unfolded*, *Intermediate I* and *Intermediate II*. The 'three-state' folding pathway was obtained via the route containing the *Unfolded*, *Intermediate II* and *Folded* ensemble. The biphasic nature of EnHD was observed, as the HTH motif formation occurred in the fast phase followed by the slow phase involving the docking of Helix I and tertiary packing. The HTH motif is essential and acts as an intermediate in the EnHD folding. The use of advanced MD technique like REMD helped in understanding the folding of EnHD by exploring large conformational space, as well as reaching a long time scale of microseconds. In this study a near-native structure of EnHD was obtained. However, longer simulations matching the experimental folding time scales could have helped to reach the native ensemble. The results of the simulations performed are in good agreement with the experimental findings of EnHD, thus this approach can be further extended to study the folding mechanism of various fast folding proteins as well as misfolded proteins.

Acknowledgement

The authors gratefully acknowledge the Department of Information Technology (DIT), Government of India, New Delhi for

providing us with financial support. This work was performed using the "Bioinformatics Resources and Applications Facility (BRAF)" funded by DIT, New Delhi.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2010.09.007.

References

[1] C.B. Anfinsen, Principles that govern the folding of protein chains, Science 181 (1973) 223–230.
[2] C. Levinthal, Are there pathways for protein folding? J. Chim. Phys. 65 (1968) 44–45.
[3] P.E. Leopold, M. Montal, J.N. Onuchic, Protein folding funnels: a kinetic approach to the sequence–structure relationship, Proc. Natl. Acad. Sci. 89 (1992) 8721–8725.
[4] I.A. Hubner, E.J. Deeds, E.I. Shakhnovick, Understanding ensemble protein folding at atomic detail, Proc. Natl. Acad. Sci. 103 (2006) 17747–177752.
[5] M. Karplus, J. Kuriyan, Molecular dynamics and protein function, Proc. Natl. Acad. Sci. 102 (2005) 6679–6685.
[6] K.A. Dill, S.B. Ozkan, T.R. Weikl, J.D. Chodera, V.A. Voelz, The protein folding problem: when will it be solved? Curr. Opin. Struct. Biol. (2007) 342–346.
[7] J.N. Onuchic, P.G. Wolynes, Theory of protein folding, J. Mol. Graphics Modell. 19 (2001) 146–149.
[8] A. Fresht, V. Daggett, The present view of the mechanism of protein folding, Nat. Rev. Mol. Cell Biol. 4 (2004).
[9] C.M. Dobson, M. Karplus, The fundamentals of protein folding: bringing together theory experiment, Curr. Opin. Struct. Biol. 9 (1999) 92–101.
[10] K.A. Dill, H.S. Chan, From Levinthal to pathways to funnels, Nat. Struct. Biol. 4 (1997) 10–19.
[11] M. Karplus, The Levinthal paradox: yesterday and today, Fold Des. 2 (1997) S69–S72.
[12] C.M. Dobson, A. Sali, M. Karplus, Protein folding: a perspective from theory and experiment, Angew. Chem. Int. Ed. 37 (1998) 868–893.
[13] T.L. Religa, C.M. Johnson, M.V. Dung, S.H. Brewer, R.B. Dyer, A.R. Fresht, The helix-turn-helix as an ultrafast independently folding domain: the pathway of folding of Engrailed Homeodomain, Proc. Natl. Acad. Sci. 104 (2007) 9272–9277.
[14] F. Huang, G. Settanni, A. Fresht, Fluorescence resonance energy transfer analysis of the folding pathway of Engrailed Homeodomain, Protein Eng. Des. Sel. 21 (2008) 131–146.
[15] U. Mayor, C.M. Johnson, V. Daggett, A. Fresht, Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation, Proc. Natl. Acad. Sci. 97 (2000) 13518–13522.
[16] S. Gianni, N.R. Guydosh, F. Khan, T.D. Caldas, U. Mayor, G.W.N. White, M.L. DeMarco, V. Daggett, A.R. Fresht, Unifying features in protein-folding mechanisms, Proc. Natl. Acad. Sci. 100 (2003) 13286–13291.
[17] T. Ternstrom, U. Mayor, M. Akke, M. Oliveberg, From snapshot to movie: φ analysis of protein folding transition states taken one step further, Proc. Natl. Acad. Sci. 96 (1999) 14854–14859.
[18] M. Shen, K. Freed, All-atom Fast protein folding simulations: the villin headpiece, Proteins Struct. Funct. Genet. 49 (2002) 439.
[19] H. Lei, Y. Duan, Two-stage folding of HP-35 from ab initio simulations, J. Mol. Biol. 370 (June (1)) (2007) 196–206.
[20] H. Lei, X. Deng, Z. Wang, Y. Duan, The fast-folding HP35 double mutant have a substantially reduced primary folding free energy barrier, J. Chem. Phys. 129 (October (15)) (2008) 155104.
[21] H. Lei, Y. Duan, Ab initio folding of albumin binding domain from all-atom molecular dynamics simulation, J. Phys. Chem. B 111 (May (19)) (2007) 5458–5463.
[22] H. Lei, C. Wu, H. Liu, Y. Duan, Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations, Proc. Natl. Acad. Sci. U.S.A. 104 (March (12)) (2007) 4925–4930.
[23] Y. Chebaro, X. Dong, R. Laghaei, P. Derreumaux, N. Mousseau, Replica exchange molecular dynamics simulations of coarse-grained proteins in implicit solvent, J. Phys. Chem. B 113 (January (1)) (2009) 267–274.
[24] D.A. Beck, G.W. White, V. Daggett, Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations, J. Struct. Biol. 157 (3) (2007) 514–523.
[25] U. Mayor, N.R. Guydosh, C.M. Johnson, J.G. Grossmann, S. Sato, G.S. Jas, S.M.V. Freund, D.O.V. Alonso, V. Dagett, A.R. Fresht, The complete folding pathway of a protein from nanoseconds to microseconds, Nature 420 (2003).
[26] Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding, Chem. Phys. Lett. 314 (1999) 141–151.
[27] N.D. Clarke, C.R. Kissinger, J. Desjarlais, G.L. Gilliland, C.O. Pabo, Structural studies of the Engrailed Homeodomain, Protein Sci. 3 (1994) 1779–1787.
[28] M.L. DeMarco, D.O.V. Alonso, V. Daggett, Diffusing and colliding: the atomic level folding/unfolding pathway of a small helical protein, J. Mol. Biol. 341 (2004) 1109–1124.
[29] D. Li, Y. Haijun, L. Han, S. Huo, Predicting the folding pathway of Engrailed Homeodomain with a probabilistic roadmap enhanced reaction-path algorithm, Biophys. J. 94 (2008) 1622–1629.
[30] V. Daggett, Molecular dynamics simulations of the protein unfolding/folding reaction, Acc. Chem. Res. 35 (2001) 422–429.
[31] M.E. McCully, D.A.C. Beck, V. Daggett, Microscopic reversibility of protein folding in molecular dynamics simulations of the Engrailed Homeodomain, Biochemistry 47 (2008) 7079–7089.
[32] R. Day, V. Daggett, Direct observation of microscopic reversibility in single-molecule protein folding, J. Mol. Biol. 366 (2) (2007) 677–686.
[33] D.A. Case, T.E. Cheatham III, T. Darden, H. Gohlke, R. Luo Jr., K.M. Merz, A. Onufriev, C. Simmerling, B. Wang, R. Woods, The Amber biomolecular simulation programs, J. Comput. Chem. 26 (2005) 1668–1688.
[34] G.G. Maisuradze, A. Liwo, H.A. Scheraga, Principle component analysis for protein folding dynamics, J. Mol. Biol. 385 (2009) 312–329.
[35] S.J. Hubbard, J.M. Thornton, NACCESS, Department of Biochemistry and Molecular Biology, University College London, 1993, NACCESS.
[36] M. Heinig, D. Frishman, STRIDE: a Web server for secondary structure assignment from known atomic coordinates of proteins, Nucl. Acids Res. 32 (2004), W500-2.
[37] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics. J. Mol. Graphics (1) (1996) 33-8, 27-8.