Application of principal component analysis in protein unfolding: An all-atom molecular dynamics simulation study

Atanu Das and Chaitali Mukhopadhyay

Citation: The Journal of Chemical Physics **127**, 165103 (2007); doi: 10.1063/1.2796165
View online: http://dx.doi.org/10.1063/1.2796165
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/127/16?ver=pdfcov
Published by the AIP Publishing

Application of principal component analysis in protein unfolding: An all-atom molecular dynamics simulation study

Atanu Das and Chaitali Mukhopadhyay[a)]
Department of Chemistry, University of Calcutta, 92 A.P.C. Road, Kolkata 700009, India

(Received 20 August 2007; accepted 18 September 2007; published online 29 October 2007)

We have performed molecular dynamics (MD) simulation of the thermal denaturation of one protein and one peptide—ubiquitin and melittin. To identify the correlation in dynamics among various secondary structural fragments and also the individual contribution of different residues towards thermal unfolding, principal component analysis method was applied in order to give a new insight to protein dynamics by analyzing the contribution of coefficients of principal components. The cross-correlation matrix obtained from MD simulation trajectory provided important information regarding the anisotropy of backbone dynamics that leads to unfolding. Unfolding of ubiquitin was found to be a three-state process, while that of melittin, though smaller and mostly helical, is more complicated. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2796165]

I. INTRODUCTION

Folding of proteins is a complex phenomenon. Specific three-dimensional structures of proteins originate via a myriad of conformational changes. According to classical protein folding theory, the process occurs via the formation of a nearly sequential series of discrete intermediates.[1–3] However, the energy landscape theory describes the phenomenon as sequential progressive organization of an assembly of partially folded structures through which protein gets its natively folded structure.[4–10] Specific three-dimensional protein structure evolves from the heterogeneity of the protein chain. Relative orientations and positioning of the residues lead to a difference in the energy distribution and this, in turn, enable some folded structures to be more stable than others.

Characterization of the pathway of protein unfolding process is one of the most popular targets in modern science. Molecular dynamics (MD) simulation, coupled with experimental results, has been extensively used to gain insight into the problem. Three main techniques have been applied so far, which are denaturation by thermal/mechanical perturbation, biased-sampling free-energy method,[11–13] and targeted molecular dynamics.[14,15] The main problem with MD simulation is the time needed to study protein dynamics. The accessible simulation times on the order of hundreds of nanoseconds are much shorter than the micro to millisecond time scales at which folding/unfolding of proteins occur. Several attempts have been made to increase the molecular dynamics simulation time scale.[16–18]

Our work is based on the use of principal component analysis (PCA),[19] which is also known as Karhunen-Loeve expansion. It is a technique to reduce the number of degrees of freedom. Karplus and Kushick introduced this method,[20,21] under the name "quasiharmonic analysis," to the protein research community. By this method, molecular dynamics trajectories can be explained satisfactorily in terms of small number of variables (essential degrees of freedom). Interestingly, less than 5% of all degrees of freedom can describe more than 90% of the total atomic motion by PCA or singular value decomposition.[19,22–24] The residues whose contributions are important in the formation and stability of different structural elements can be identified by PCA. The covariance matrix,[19,22,25–28] whose diagonalization gives PCA, is important in describing the pathway of the dynamic process and the relation between the movements of different regions of the molecule during the process. Several works have already been performed utilizing PCA to identify protein dynamics.[29–33] The objective of our work is to identify the following:

(i) All the relevant structural changes leading to unfolding at atomic resolution;
(ii) contribution of the individual residues towards the global dynamics;
(iii) correlation/anisotropy in the dynamics leading to local disruption of native contacts; and
(iv) the weak native contacts which are crucial to maintain the three-dimensional structure.

We have applied the technique here to two biologically important molecules—(1) ubiquitin and (2) melittin.

Ubiquitin, a small globular protein, has been monitored extensively as a model system for investigating the factors governing the stability of proteins for the past twenty years. In recent years, the kinetic model of folding for ubiquitin has become somewhat controversial as both the two-state (native–unfolded) and three-state (native–intermediate–unfolded) models of folding have been reported.[34–38] Many papers have been published on the possibility of transition of folding pathway from two-state to three-state fashion, which might be due to the effect of stabilizing salts[39] or low temperature.[40] We had shown earlier[41] that, during the unfolding of ubiquitin, two transition states are visited: One resembles the well known "A" state and the other containing

[a)]Electronic mail: chaitalicu@yahoo.com

only the N-terminal β-hairpin and a part of the α-helix. In our present study, we have given a complete picture of the stability order of the different structural fragments of the protein along with their contribution to the unfolding dynamics. We have also answered whether the motion of different structural parts of the protein are correlated or not. We have characterized the local fluctuations that lead to the destabilization of the three-dimensional arrangements leading to unfolding of the protein. We have shown that unfolding, and hence folding, of ubiquitin is a three-state process which has also been shown recently from mass spectroscopic analysis.[42] The unfolding potential energy funnel shows that there is no reverse energy barrier between the so-called "A state" and unfolded state of ubiquitin.

In the membrane-bound bioactive form, melittin adopts an alpha-helical conformation with two helices, which are at an angle of 120° to each other.[43] Transition of the structure of melittin from random coil to helix is of immense importance in biological sciences. To investigate the contribution of each of the residues towards the stability of the helical form and to identify whether the helix-coil transition is a sequential or parallel process, we have performed thermal unfolding of melittin monomer at 450 K (near T_m) using all-atom molecular dynamics simulation. Melittin is one of the shortest peptides having only α-helices as its secondary structure.

II. METHODS

All molecular dynamics simulations were performed using CHARMM22 force field and parameters. The native structure of ubiquitin was obtained from protein data bank (PDB code: 1UBQ). For simulation run, the starting structure of the protein was immersed in a cubic box ($52\times52\times52$ Å^3) of TIP3P water molecules, so that all water molecules with oxygen atoms less than 2.6 Å from the protein were removed. The system was energy minimized using the steepest descent method until the difference between the total potential energies of the molecular system at consecutive energy minimization steps reaches a value of less than 0.001 kJ mol^{-1}. A series of three simulations was carried out in a periodic box of water with 2882 water molecules in the *NPT* ensemble (constant number of particles, pressure, and temperature) at $T=373$ K and $P=1$ atm. An 8 Å cutoff was applied to nonbonded interactions, and the particle-mesh Ewald summation was applied to evaluate electrostatic interactions. The SHAKE algorithm was used to maintain the valence geometry of the water molecules. The simulations were carried out for 18 ns and the integration time step was 2 fs. For melittin, the "A" chain of the dimer obtained from the protein data bank (PDB code: 2MLT) was dissolved in a orthorhombic box ($48\times40\times32$ Å^3) of 1531 water molecules of TIP3P water model. The rest of the simulation protocols are the same as mentioned above for ubiquitin. Three simulations were carried out for 6.5 ns each at 450 K.

Covariance matrix was calculated for each of the two systems. For this analysis, only the backbone C_α atoms have been considered. Each covariance matrix represents the correlation among the residues averaged over a particular trajectory (18 ns for ubiquitin and 6.5 ns for melittin). This is based on the following expression:

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle, \tag{1}$$

where $\langle\ \rangle$ is the average over all data points. The protein moves in a N-dimensional space, where N is three times the number of atoms. Only a few degrees of freedom will contribute significantly to the global fluctuation of the protein. Covariance or principal component has been used to find these degrees of freedom. PCA can be obtained from covariance matrix by choosing an orthonormal matrix which diagonalizes C_{ij}. Then the above equation can be written as

$$C_{ij} = R\ \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N) R^T. \tag{2}$$

The ith column of the matrix R is the eigenvector of principal mode corresponding to eigenvalue λ_i. The eigenvalue is the mean square fluctuation in the direction of the principal mode. Contributions of individual C_α atoms towards a particular principal component were calculated. As this contribution in principal components is the result of the fluctuation of each C_α atom in the x, y, and z directions (in Cartesian coordinate), fluctuations of each C_α atom were resolved along three orthogonal directions (x,y,z) with respect to a common frame. Maximum displacement of a particular C_α atom is then plotted for first four PCs for ubiquitin and first three PCs for melittin.

To draw the potential energy surface of the proteins, we have calculated the root mean square deviation (RMSD), R_g, and potential energy of the system as a function of time for all the three simulations of ubiquitin and melittin. Thus, for each combinations of RMSD and R_g experienced by the protein, the potential energy values were assembled. We have then used the ORIGIN program (version 6.0) to generate the three-dimensional data grid using correlation method. The correlation method computes a new value for each cell in the matrix from the values of the points in the adjoining cells in the matrix that are included within the search radius. Thus smoothness of the surface is achieved.

III. RESULTS

Results obtained from three different simulation trajectories are in good agreement for the respective systems. So we have presented the analysis obtained from a single trajectory.

A. Ubiquitin

Ubiquitin is a small 76-residue globular protein. The melting temperature of ubiquitin is 373 K. In its bioactive form it has five beta strands—$\beta1(1-7)$, $\beta2(10-17)$, $\beta3(40-45)$, $\beta4(48-50)$, and $\beta5(64-72)$; and two different kinds of helices—$\alpha1(23-34)$ and 3_{10} helix(56–59)—in its bioactive form.[44–46] A total of 54 ns of unfolding simulations with explicit solvents have been recorded. We describe here the independent as well as correlated inter-residue motions based on cross-correlation matrix and the principal components obtained from them.[47–49]

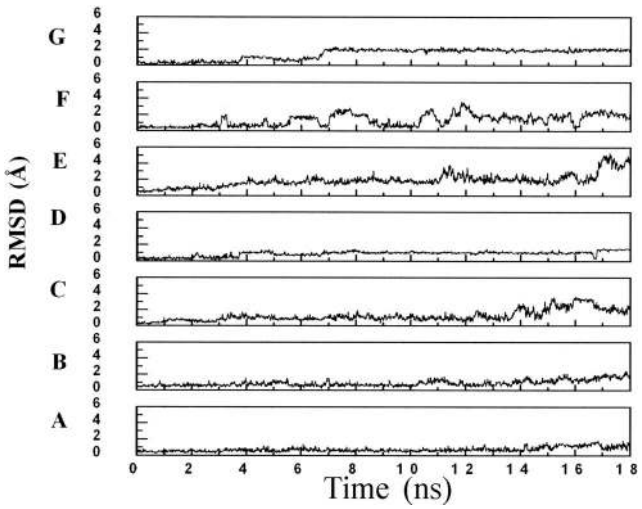

FIG. 1. Root mean square deviation of different structural fragments of ubiquitin at 373 K—from (A) to (G) we have RMSD values obtained in the 18 ns simulation of ubiquitin—(A) β1, (B) β2, (C) β3, (D) β4, (E) β5, (F) α-helix, and (G) 3_{10} helix.

1. Root mean square deviation

To understand the unfolding process of Ubiquitin, backbone RMSD of each secondary structural element of the protein (β1, β2, β3, β4, β5, α1, and 3_{10}) was calculated as compared to the native structure and plotted as a function of time in Fig. 1. RMSDs of β1 and β2 retain a value of less than 1 Å until towards the end where both increase. However, RMSD of β3 fluctuates near a constant value of 1.0 Å starting from 0.3 Å, and after 14 ns of simulation, it reaches a value close to 3.5 Å. Finally, it gets a value of 2.0 Å at the end of simulation. RMSD of β5 increases from 0.5 to 2.0 Å within 4 ns of simulation and fluctuates almost around it up to 15 ns. Then a sudden increase is observed and the value goes to near 4.5 Å. Up to 5 ns, RMSD of α1 remains near 0.5 Å, and after that, large fluctuations are observed (maximum value of 3.5 Å at 12 ns). The feature is somewhat different for 3_{10} helix, where RMSD is nearly constant with a value of 0.5 Å up to 4 ns and then it increases suddenly to 1.0 Å. Another sharp jump in RMSD is observed just after 6 ns of simulation when the value of RMSD increases from 0.6 to 2.0 Å and remains constant till the end of simulation. The course of events leading to the observed RMSD fluctuations is discussed later.

2. Hydrogen bond framework

All the secondary structural elements of ubiquitin are connected to each other by a large number of hydrogen bonds. Rigid hydrogen bond framework plays a key role in the stability of the protein. Duration of existence of each hydrogen bond in the simulation trajectory is summarized in Table I (H bonds between two secondary structures) and Table II (H bonds between loop regions with other secondary structures or loops). We had observed similar trends in the loss of hydrogen bond pattern in all the trajectories, though the exact time points are slightly different. The sequence of loss of different interfaces is apparent from these two tables. An important aspect of ubiquitin unfolding is the dynamics

TABLE I. Survival times of the H bonds that are between different secondary structural fragments of ubiquitin. (SC denotes side chain.)

Contacts	H bond	Disappearance (ns)	Reappearance (ns)
β1-β2	M:1:SC–V:17:O	13.81	
	M:1:O–V:17:N	13.85	
	I:3:N–L:15:O	14.18	
	I:3:O–L:15:N	Stable	
	V:5:N–I:13:O	Stable	
	V:5:O–I:13:N	Stable	
	T:7:N–K:11:O	10.26	11.60 (fluctuation)
	T:7:SC–K:11:N	10.26	11.60 (fluctuation)
β4-β4	R:42:SC–Q:49:SC	0.69	
	R:42:SC–Q:49:SC	0.67	
	L:43:O–L:50:N	0.92	
	F:45:N–K:48:O	1.91	
	F:45:O–K:48:N	0.8	Near 17 ns for 80 ps
β1-β5	Q:2:O–E:64:N	3.26	
	F:4:N–S:65:O	8.04	
	K:6:O–L:69:N	15.5	
	K:6:N–L:67:O	16.8	
	F:4:O–L:67:N	16.78	
β3-β5	R:42:N–V:70:O	3.36 and 10.44	7.2
	Q:40:O–R:72:N	3.08 and 10.32	7.2
	I:44:O–H:68:N	7.24	
	I:44:N–H:68:O	7.92	
	R:42:O–V:70:N	10.44	
α1-β2	E:34:SC–K:11:SC	2.7 and 6.04	4.04
	E:34:SC–K:11:SC	2.7 and 6.04	4.04
α1-β3	K:27:O–Q:41:SC	2.15	

of the two helices. Time series of the number of the characteristic H bonds of the two helices at different time points show that the i-(i+4) H bond which is the characteristics of α-helix remains near a constant value of 6 fluctuating with a deviation of ±2. But just after 7 ns of simulation, its value decreases suddenly to a value of 3. After 8 ns, it again increases and then a gradual decrease in number of H bonds is observed. But for the 3_{10} helix, the number of H bonds involved in helix packing fluctuates from 0 to 2. The α-helix has retained some residual helical nature even at the end of the trajectory.

3. Covariance analysis

The covariance matrix for C_α position fluctuations of ubiquitin at 373 K with explicit water model is shown in Fig. 2. This includes both through-space and contact interactions. The values of correlation are indicated at the right side of the plot. Extent of correlations can be classified in three ranges—(1) 1–0.75, (2) 0.75–0.50, and (3) 0.50–0.25 (called as high, moderate, and weak correlations, respectively). The residues of β1 and β2 are highly correlated with each other and also with the residues of β5. The terminal residues of β5 are correlated with the β3 and β4 but the values are smaller. The white regions indicate low correlation. As expected, all the portions of ubiquitin are not strongly correlated with each

TABLE II. Survival times of the H bonds that are not only between different secondary structural fragments of ubiquitin. (SC denotes side chain.)

	H bond	Disappearance (ns)	Reappearance (ns)
Hydrogen bonds that are not between two secondary structural elements of ubiquitin	T:14:SC–T:14:O	Stable	
	D:21:O–L:56:N	1.67	
	T:22:N–N:25:SC	2.02 and 13.53	11.32
	T:22:O–N:25:N	13.88	
	I:23:N–R:54:O	1.87	
	K:27:SC–R:54:SC	1.69	
	I:36:O–Q:41:SC	2.92	
	P:37:O–Q:40:N	3.26 and 11.21	Near 7 and 10
	P:38:O–Q:41:N	0.99	
	D:39:O–R:72:SC	2.81 and 8.98	8.54
	E:51:SC–R:54:SC	3.12	
	E:51:SC–R:54:SC	3.12	
	T:55:N–D:58:SC	3.33	
	T:55:O–D:58:N	6.01	
	S:57:O–N:60:SC	From the beginning	
	K:63:SC–E:64:SC	Fluctuating	
	K:63:O–S:65:N	3.00	
	R:72:SC–R:74:O	15.34 (fluctuates)	
	R:72:SC–R:74:O	15.34 (fluctuates)	
	R:74:SC–G:76:SC	Fluctuating	
	R:74:SC–G:76:SC	Fluctuating	

other. $\alpha 1$ of ubiquitin is not correlated with any other part of the protein and this fact has important consequence on the ubiquitin unfolding as discussed later.

4. Principal component analysis

To analyze the unfolding dynamics of the protein, PCs have been considered (leaving the first six components for translational and rotational degrees). More than 90% of the atomic motions are included in these principal components (figure not shown). The plot of PCs as a function of time (figure not shown) indicates that PC1, PC2, and PC4 contribute from the beginning of the dynamics and PC3 contributes towards the end. No large fluctuation was observed for PC5 and was not considered for further analysis. Contributions of individual C_α atoms towards a particular principal component are obtained after the diagonalization of the covariance matrix. The diagonal values contain components of the displacement of the individual atoms along three Cartesian axes. The coefficients corresponding to PC1 to PC4 for individual residues of the protein were calculated and resolved in three orthogonal directions. The maximum values of the coefficients along any of the three directions were plotted against the residue number (figure not shown).

In case of PC1, the β sheets mostly contribute with $\beta 1$, $\beta 2$, $\beta 3$, and $\beta 5$ having $+ve$ displacement along a particular direction, whereas $\beta 4$ moves in opposite direction along a different axis. The residues of C terminus show alternative $+ve$ and $-ve$ bars of high magnitude of coefficients indicating disruption of the $\beta 3/\beta 4$ and $\beta 4/\beta 5$ interfaces. PC2 gets contribution from the α-helix movement as well as very anisotropic movement of the C terminus. The contribution of PC3 increases drastically after 14 ns and we can see highly anisotropic dynamics in the residues from 20 to 76. The movements of $\beta 1$ and $\beta 2$ are also along the opposite directions. Interestingly, this direction gets reversed in PC4 leaving the $\beta 1/\beta 2$ interface more or less intact.

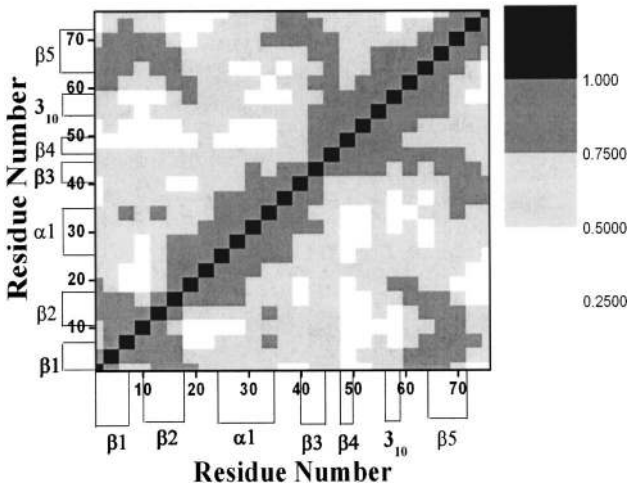

FIG. 2. Correlation matrix of ubiquitin obtained at 373 K averaged over total simulation time. Cross-correlation matrix or covariance of C_α atoms of ubiquitin calculated from explicit water simulation, value of correlation between the C_α atoms decreases from black to white, and different ranges of the values are indicated in the adjacent column. The secondary structures are marked along the two axes.

B. Melittin

Melittin is a 26-residue amphipathic peptide. The melting point of melittin is 448–455 K (175–182 °C). In its bioactive form, it has two helices and a turn as its secondary structural counterparts. The helices are from residue numbers 1–11 and 14–26.[43,47–49] Unfolding of melittin was studied at three different temperatures—300, 450, and 550 K by MD simulation. For analysis, only the unfolding trajectories at 450 K (simulation time of 6.5 ns) has been taken into account. At 300 K, the peptide chain does not completely unfold for several nanoseconds. At 550 K, the unfolding is very rapid (unfolds within 1 ns); as a result, the gradual change in the native structure of the polypeptide cannot be analyzed clearly. The trajectory at 450 K is a good one to study the unfolding process as it is also the transition temperature of melittin. Three independent trajectories with explicit water model (each of 6.5 ns) were obtained.

1. Root mean square deviation

RMSDs of $\alpha 1$, $\alpha 2$, and turn were plotted against simulation time (Fig. 3). RMSD of $\alpha 2$ fluctuates from 0.25 to 2.25 Å almost up to 2 ns. Within this period of time, RMSD of the turn is near 1.0 Å with occasional fluctuation, and after that, it decreases from 1.5 to 0.5 Å. RMSD of $\alpha 1$ lies near 3.0 Å. However, after 2 ns, RMSD of $\alpha 2$ suddenly increases from a value near 1.0 to 3.0 Å and then it increases gradually to almost 4.0 Å.

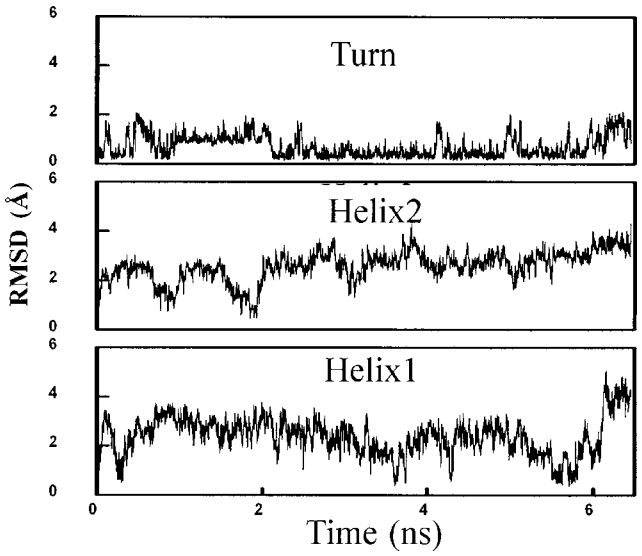

FIG. 3. Root mean square deviation of three important structural fragments of melittin as a function of simulation time.

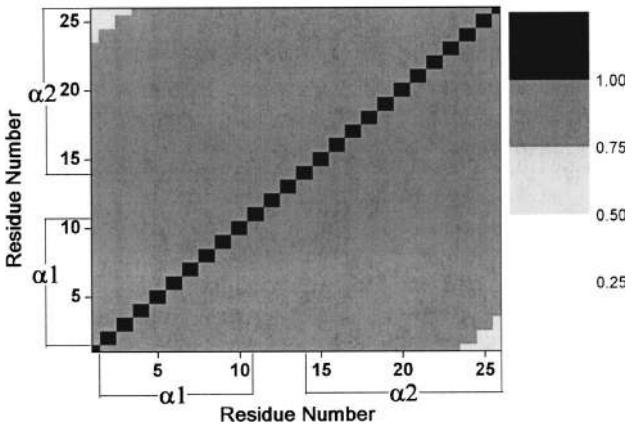

FIG. 4. Correlation matrix of melittin obtained at 450 K averaged over total simulation time. Cross-correlation matrix or covariance of C_α atoms of melittin calculated from explicit water simulation, value of correlation between the C_α atoms decreases from black to white, and different ranges of the values are indicated in the adjacent column. The secondary structures are marked along the two axes.

2. Hydrogen bond framework

The survival time of i-$(i+4)$ hydrogen bond was calculated for both helices of melittin (Table III). It fluctuates rapidly for $\alpha 1$ with time. The number of those hydrogen bonds lies within 1–4 for $\alpha 1$ and finally it becomes zero. But for $\alpha 2$, it lies within 4–8 which decreases rapidly after 2 ns of simulation. For both the helices, hydrogen bonds show a fluctuating nature from 4 to 6 ns.

3. Covariance analysis

Melittin took about 6.5 ns to unfold completely at 450 K. We have calculated covariance matrix where information is averaged over total simulation time (Fig. 4). The motions of the residues of $\alpha 1(1-11)$ are highly correlated having magnitude in the range of 0.80–0.99, whereas other parts of melittin (mainly the residues of the $\alpha 2$) are correlated with the helix1 by a value of 0.60–0.80.

TABLE III. Survival times of the H bonds of melittin.

	H bonds	Disappearance (ns)	Reappearance (ns)
Hydrogen bonds between the residues of melittin	G:1:O–V:5:N	0.33, 3.75 and 5.87	3.44, 5.23
	I:2:O–L:6:N	0.32, 3.76 and 5.88	3.15, 5.23
	G:3:O–K:7:N	0.44 and 6.14	1.55
	A:4:O–V:8:N	0.44 and 6.13	1.17
	V:5:O–L:9:N	6.11	
	L:6:O–T:10:N	6.11	
	V:8:O–TG12:N	5.91	
	V:8:O–L:13:N	0.88 and 5.96	2.00
	G:12:O–L:16:N	Fluctuating	
	L:13:O–I:17:N	Fluctuating	
	P:14:O–S:18:N	Fluctuating	
	A:15:O–W:19:N	Fluctuating	
	L:16:O–I:20:N	Fluctuating	
	I:17:O–K:21:N	5.64	
	S:18:O–R:22:N	2.63 and 5.79	3.02
	W:19:O–K:23:N	2.53 and 2.51	3.02
	I:20:O–R:24:N	2.22 and 5.51	5.03
	K:21:O–Q:25:N	2.14	
	R:22:O–Q:26:N	2.06	

4. Principal component analysis

It was observed that three principal components contribute nearly 90% of the total dynamics. The plot of three principal components against simulation time (figure not shown) indicates that PC1 and PC2 have the maximum contribution towards unfolding. Initial dynamics was governed by PC1 and, after that, PC2 adds to it. PC3 has little effect at the beginning, but the extent increases towards later part of the trajectory. As has been done for ubiquitin, the coefficients of PC1, PC2, and PC3 in the three perpendicular directions were calculated for all of the residues of melittin, and the maximum of the values for each residue corresponding to each principal component were plotted (figure not shown). We see from the plot that the C-terminal end of $\alpha 1$ and $\alpha 2$ show a twisted motion. In PC3, the turn (residue number 13 and 14) shows a correlated motion, but in PC1, it shows a twisted motion.

IV. DISCUSSION

A. Unfolding pathway of ubiquitin

It is well established that ubiquitin, being a very small single domain protein, has two distinct regions that display a dramatic difference in dynamics upon mild denaturation.[38,42] We had previously reported that the sequence of events leading to unfolding of ubiquitin under thermal denaturation proceeds via a transition state retaining the N-terminal hairpin and the α-helix.[41] The present study is more extensive. We have recorded a total of 54 ns unfolding trajectory of ubiquitin. Surprisingly, the sequence of unfolding events more or less remains the same as reported earlier,[41] which is indicative of a well-defined unfolding pathway for the protein ubiquitin.

The covariance matrix shows high level of interstrand correlation. The plot of the maximum contribution of indi-

vidual C_α atoms to a particular principal component indicates that the local dynamics is highly anisotropic. The coefficients clearly show how local displacements of the main chain can give rise to global dynamics. This has been shown recently by Horner *et al.*[42] and by Chung *et al.*[50] using two-dimensional IR spectroscopy and hydrogen exchange mass spectrometry, respectively. Both these experimental techniques have suggested C-terminal end to be more dynamic than the N-terminal end, and we report here the exact contacts that are lost during various points of the simulation. At the initial stages of the unfolding, PC1 and PC2 contribute significantly and one can see the uniform movement of the N-terminal segment while that of the C-terminal segment is anisotropic. Early opening of the $\beta3/\beta4$ interface is also indicated in Table I, where all the hydrogen bonds connecting the two sheets disappear rapidly. Opening of the $\beta3/\beta4$ interface is accompanied by the rupture of the $\beta3/\beta5$ interface. In PC2, the residues of C-terminal β sheets show movements in different directions leading to loss of contacts. This indicates that the thermal perturbation allows amplified configurational flexibility in the more weakly interacting strands $\beta3-\beta5$, which is similar to the events reported by Chung *et al.*[50] Horner *et al.*[42] showed that in the "A" state $Gly^{53}-Gly^{76}$ has the highest degree of structural disorder, while the segment $Thr^{9}-Glu^{18}$ shows the least. In PC1 and PC3, we indeed observe highly anisotropic movements in the C terminus, while the N-terminal segment moves in a concerted manner.

After opening of the C-terminal portion of the protein, dynamics move towards the N terminus. Initiation of denaturation of N-terminal part of the protein is demonstrated in PC3 and PC4. The most important feature of PC3 is the dynamics of the α-helix. In PC1 and PC2, the helical residues show correlated motions, but the exception occurs at PC3. All the residues of the α-helix fluctuate in different reference axes and in different directions, i.e., their movements are no longer isotropic. This leads to the partial unfolding of the α-helix. But complete unfolding of α-helix is not observed in the simulation which is supported by the presence of residual helical nature of the residues even at the end of the simulation. This is in agreement with the hydrogen exchange mass spectrometry analysis by Horner *et al.*,[42] where a partial protection in $Pro^{19}-Gln^{34}$ segment was observed that contains the α-helix. The contribution of PC4 increases after 12 ns of simulation, showing large scale movement of almost the entire chain. In PC3, the $\beta1$ and $\beta2$ strands have maximum amplitude motion along the same axis but in opposite direction which gets reversed in PC4. This indicates that, though there is large amplitude motion in $\beta1$ and $\beta2$ residues, the hairpin structure is retained. The $\beta1/\beta2$ interface does not completely open up to a random coil structure due the three stable native hydrogen bonds—one between I3-L15 and the other two between V5-I13.

"The rise in thermal energy in the solvent and protein allows increased configurational flexibility in the more weakly interacting and partially denatured strands III–V. In the second stage, the remaining hairpin involving the more stable strands I and II unfolds."[50] We also observe here that longer time scale is needed for N-terminal β hairpin and the α-helix to unfold, which is in good agreement with experimental data reported by Chung *et al.*[50] C-terminal segment of the protein exhibits faster motional character, while the N-terminal segment shows slow movements which supports the phenomenon presented by Horner *et al.*[42] Amazingly, irregular allocation of the backbone elasticity is observed in the non-native state of ubiquitin. This may be a significant structural characteristic.[42,51]

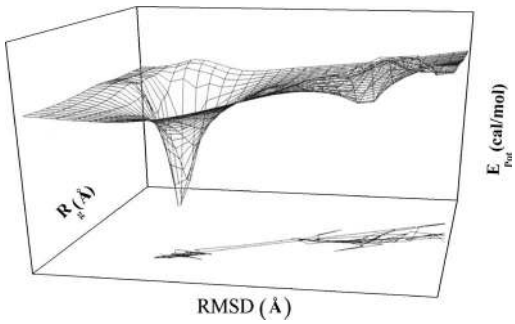

FIG. 5. Potential energy landscape of ubiquitin (three state)—three-dimensional (3D) representation. At the bottom of the 3D diagram, motion of the Ubiquitin molecule in one of the three trajectories is projected as a function of RMSD and R_g.

B. Transition state ensemble of ubiquitin

Pathways of protein folding can be described by free-energy diagrams. Usually, fraction of native contacts and radius of gyration are chosen as reaction coordinates. In the present case, the potential energy of the protein at different points of simulation is plotted (as discussed in Sec. II) as a function of RMSD and radius of gyration (R_g), both of which are loosely defined folding reaction coordinates (Fig. 5). Here, we have three basins of importance—native (N), intermediate (A), and unfolded (U)—in this potential energy surface. N state is represented by a sharp potential well; A state is represented by a comparatively diffused and higher-energy local minimum; whereas the U state cooperatively defines all the other conformations, i.e., a collection of several microstates. There is a striking similarity between Fig. 5 and the free-energy surface obtained by Qin *et al.*[40] and Horner *et al.*[42] It is seen from the plot that the reverse activation energy from U state to A state is negligible (almost nonexistent). As a result, the possibility of having a particle in the U state is very small as all tend to go to the lower-energy A state which is almost identical to the conclusions drawn by Horner.[42] The unfolding of ubiquitin or the reverse process will then be essentially a three-state (i.e., native $\rightarrow$intermediate$\rightarrow$unfolded) one. The transition state ensemble of the protein has two different ranges of R_g values—one from 12 to 13 Å and the other from 13 to 14 Å (obtained from the distribution curve of the population of states in different ranges of R_g). These correspond to two types of conformations—one with almost native-like character having four beta sheets along with a partially denatured $\beta3/\beta4$ hairpin and α-helix, while the other consists of a partially unfolded $\beta1/\beta2$ hairpin and some residual helical nature in α-helix. All the above findings closely resemble the structures reported using φ, ψ analysis by Sosnick *et al.*[52] The ΔE

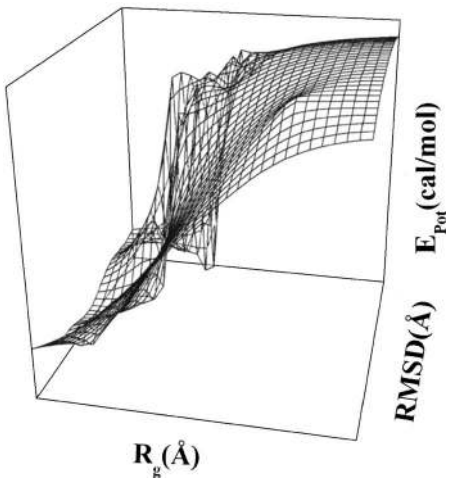

FIG. 6. Potential energy landscape of melittin (multistate)—three-dimensional representation.

between N and A states, obtained from Fig. 5 is 36.42 kJ/mol, which is very close to the value (35.00 kJ/mol) suggested by Qin *et al.*,[40] from their experimental observations.

C. Unfolding pathway of melittin

Two helices in melittin are different in order of their stability as well as functions. The residues K7, V8, L9, and T10 of $\alpha 1$ first participate in unfolding via a twisted motion which is observed in PC1. The dynamics then moves to the N terminus of $\alpha 1$ and the same kind of twisted motion is observed in G3, A4, and V5 indicated in PC2. This kind of motion drives the helical residues away from each other and the hydrogen bond framework of $\alpha 1$ starts to break down. Residues of $\alpha 1$ are only weakly correlated with the residues of $\alpha 2$ (Fig. 4). As a result, the two helices unfold rather independently. Thermal perturbation is carried to $\alpha 2$ through the interhelix turn, mainly by L13 and P14, due to the twisted movement of these residues indicated in PC1. Initiation of unfolding in $\alpha 2$ occurs at K21, R22, and R24. After 4 ns, the high correlated movement of the residues of $\alpha 2$ breaks the rigid H-bond framework of the helical structure (Table III). Then the residual helical nature of $\alpha 1$ vanishes (evident from PC3 which exhibits twisted motion at those residues) and melittin unfolds to a complete random coil structure. So, unfolding of the turn triggers the opening of $\alpha 2$. During the unfolding pathway, melittin undergoes a series of fluctuations. A particular part of the peptide, which loses initially the secondary structural features, may suddenly gain it at any position of the process as is observed for $\alpha 1$.

D. Transition state ensemble of melittin

The same procedure is applied for melittin as done before for ubiquitin. The potential energy of the peptide is plotted as a function of RMSD and R_g (Fig. 6). Though melittin is a small peptide, its potential energy diagram is far more complex than ubiquitin. Unlike ubiquitin, unfolding of melittin or the reverse process is not a three-state one. More than one diffused low energy potential wells are observed both at low and high RMSD and R_g values, indicating higher degrees of heterogeneity compared to ubiquitin. Consequently, the transition state ensemble has no fixed conformation. The transition state conformations have wide range of R_g values—one from 9 to 10 Å, one from 10 to 11 Å, and the other from 11 to 12 Å. The ΔE between the native and lowest energy intermediate is 0.27 kJ/mol and the ΔE between from the lowest energy intermediate and unfolded state is 9.28 kJ/mol (Fig. 6).

V. CONCLUSIONS

We report here a method to describe the extent of movement of individual residues of a protein/peptide along the unfolding pathway. Calculation of coefficient of principal component can enable us to identify individual contribution behind various dynamical phenomena in a molecule. We have identified here the importance of dynamic correlations among the residues of a protein or peptide during unfolding from its native or bioactive structure to a random structure having small or no similarity with the native organization. Our work reveals both the direction and amplitude of motion of a particular residue in the unfolding pathway which is a completely new outcome of PCA. The potential energy mapping is validated by the experimental evidence of ubiquitin using mass spectroscopy.[42] Unfolding sequence of ubiquitin is different in the simulation trajectories from that obtained from mechanical perturbation,[53–55] but resembles strongly with experimental observations.[50] From the potential energy diagram, we can suggest that protein folding may be a two-state, three-state, or multi-state process, but that does not necessarily depend on the dimension and structural diversity of the protein or peptide. Ubiquitin, being a 76-residue globular protein with different types of secondary structural elements follow three-state process; whereas, for melittin with two helices, helix-coil transition takes place via a complex multistate process. The advantage of the present method is that the adopted procedures describe the process satisfactorily and enable us to have the complete detailing of the process irrespective of the pathway adopted. In both cases, the early stages of unfolding involved loss of long-range hydrogen bonds connecting residues which are part of different secondary structures/loop regions. They are the "soft" or "weak" spots of a protein structure that are likely to form at the later stages of folding. Identification of these kinds of contacts in a protein structure and replacing them with more stable contacts such as disulphide linkage or salt bridge might enhance the stability of a protein.

ACKNOWLEDGMENTS

We thank Department of Chemistry, University of Calcutta for the computational facility. One of the authors (A.D.) is thankful to CSIR, India for the financial support through CSIR-NET.

[1] P. S. Kim and R. L. Baldwin, Annu. Rev. Biochem. **51**, 459 (1982).
[2] F. X. Schmid, Biochemistry **22**, 4690 (1983).
[3] C. R. Matthews, Annu. Rev. Biochem. **62**, 653 (1993).
[4] B. D. Bursulaya and C. L. Brooks III, J. Am. Chem. Soc. **121**, 9947

(1999).
[5] B. D. Bursulaya and C. L. Brooks III, J. Phys. Chem. B **104**, 12378 (2000).
[6] P. Pokarowski, A. Kolinski, and J. Skolnick, Biophys. J. **84**, 1518 (2003).
[7] I. Tavernelli, S. Cotesta, and E. E. D. Iorio, Biophys. J. **85**, 2641 (2003).
[8] G. M. S. D. Mori, C. Micheletti, and G. Colombo, J. Phys. Chem. B **108**, 12267 (2004).
[9] A. Venkatnathan and G. A. Voth, J. Chem. Theory Comput. **1**, 36 (2005).
[10] D. J. Lacks, Biophys. J. **88**, 3494 (2005).
[11] J. E. Shea, J. N. Onuchic, and C. L. Brooks III, Proc. Natl. Acad. Sci. U.S.A. **99**, 16064 (2002).
[12] I. Daidone, A. Amadei, D. Roccatano, and A. D. Nola, Biophys. J. **85**, 2865 (2003).
[13] H. Oberhofer and C. Dellago, J. Phys. Chem. B **109**, 6902 (2005).
[14] P. Ferrara, J. Apostolakis, and A. Caflisch, J. Phys. Chem. B **104**, 4511 (2000).
[15] P. Krüger, S. Verheyden, P. J. Declerck, and Y. Engelborghs, Protein Sci. **10**, 798 (2001).
[16] V. Daggett and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **89**, 5142 (1992).
[17] Y. Duan, L. Wang, and P. A. Kollman, Proc. Natl. Acad. Sci. U.S.A. **95**, 9897 (1998).
[18] Y. Duan and P. A. Kollman, Science **23**, 740 (1998).
[19] J. Shlens, A Tutorial on Principal Component Analysis, Systems Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA and Institute for Nonlinear Science, University of California, San Diego, La Jolla, CA, 2005.
[20] M. Karplus and J. N. Kushick, Macromolecules **14**, 325 (1981).
[21] R. M. Levy, A. R. Srinivasan, W. K. Olson, and J. A. McCammon, Biopolymers **23**, 1099 (1984).
[22] I. Bahar, B. Erman, T. Haliloglu, and R. L. Jernigan, Biochemistry **36**, 13512 (1997).
[23] P. Doruker, A. R. Atilgan, and I. Bahar, Proteins: Struct., Funct., Genet. **40**, 512 (2000).
[24] S. B. Ozkan, K. A. Dill, and I. Bahar, Protein Sci. **11**, 1958 (2002).
[25] S. Swaminathan, W. E. Harte, and D. L. Beveridge, Jr., J. Am. Chem. Soc. **113**, 2717 (1991).
[26] S. F. Lienin and R. Brüschweiler, Phys. Rev. Lett. **84**, 5439 (2000).
[27] B. Hess, Phys. Rev. E **65**, 031910 (2002).
[28] B. L. Kormos, A. M. Baranger, and D. L. Beveridge, J. Am. Chem. Soc. **128**, 8992 (2006).
[29] M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten, J. Phys. Chem. **100**, 2567 (1996).
[30] A. Palazoglu, A. Gursoy, Y. Arkun, and B. Erman, J. Comput. Biol. **11**, 1149 (2004).
[31] O. F. Lange and H. Grubmuller, J. Phys. Chem. B **110**, 22842 (2006).
[32] M. Horstmann, P. Ehses, K. Schweimer, M. Steinert, T. Kamphausen, G. Fischer, J. Hacker, P. Rosch, and C. Faber, Biochemistry **45**, 12303 (2006).
[33] V. Kurkal-Siebert and J. C. Smith, J. Am. Chem. Soc. **128**, 2356 (2006).
[34] S. Khorasanizadeh, I. D. Peters, and H. Roder, Nat. Struct. Biol. **3**, 193 (1996).
[35] S. Khorasanizadeh, I. D. Peters, T. R. Butt, and H. Roder, Biochemistry **32**, 7054 (1993).
[36] B. A. Krantz and T. R. Sosnick, Biochemistry **39**, 11696 (2000).
[37] M. Schlierf, H. Li, and J. M. Fernandez, Proc. Natl. Acad. Sci. U.S.A. **101**, 7299 (2004).
[38] S. E. Jackson, Org. Biomol. Chem. **4**, 1845 (2006).
[39] H. M. Went, C. G. Benitez-Cardoza, and S. E. Jackson, FEBS Lett. **567**, 333 (2004).
[40] Z. Qin, J. Ervin, E. Larios, M. Gruebele, and H. Kihara, J. Phys. Chem. B **106**, 13040 (2002).
[41] S. G. Dasridar and C. Mukhopadhyay, Phys. Rev. E **72**, 051928 (2005).
[42] J. K. Hoerner, H. Xiao, and I. A. Kaltashov, Biochemistry **44**, 11286 (2005).
[43] T. C. Terwilliger, L. Weissman, and D. Eisenberg, Biophys. J. **37**, 353 (1982).
[44] S. Vijay-Kumar, C. E. Bugg, K. D. Wilkinson, and W. J. Cook, Proc. Natl. Acad. Sci. U.S.A. **82**, 3582 (1985).
[45] S. Vijay-Kumar, C. E. Bugg, and W. J. Cook, J. Mol. Biol. **194**, 531 (1987).
[46] S. G. Zech, A. J. Wand, and A. E. McDermott, J. Am. Chem. Soc. **127**, 8618 (2005).
[47] F. Lavialle, R. G. Adams, and I. W. Levin, Biochemistry **21**, 2305 (1982).
[48] R. Bazzo, M. J. Tappin, A. Pastore, T. S. Harvey, J. A. Carver, and I. D. Campbell, Eur. J. Biochem. **173**, 139 (1988).
[49] Y. H. Lam, S. R. Wassall, C. J. Morton, R. Smith, and F. Separovic, Biophys. J. **81**, 2752 (2001).
[50] H. S. Chung, M. Khalil, A. W. Smith, Z. Ganim, and A. Tokmakoff, Proc. Natl. Acad. Sci. U.S.A. **102**, 612 (2005).
[51] N. J. Marianayagam and S. E. Jackson, J. R. Soc., Interface **2**, 47 (2005).
[52] T. R. Sosnick, R. S. Dothager, and B. A. Krantz, Proc. Natl. Acad. Sci. U.S.A. **101**, 17377 (2004).
[53] A. Irbäck, S. Mitternacht, and S. Mohanty, Proc. Natl. Acad. Sci. U.S.A. **102**, 13427 (2005).
[54] M. S. Li, M. Kouza, and C. Hu, Biophys. J. **92**, 547 (2007).
[55] A. Kleiner and E. Shakhnovich, Biophys. J. **92**, 2054 (2007).