J. Chem. Sci. Vol. 129, No. 7, July 2017, pp. 1031–1044. © Indian Academy of Sciences.
DOI 10.1007/s12039-017-1279-1

REGULAR ARTICLE

Special Issue on THEORETICAL CHEMISTRY/CHEMICAL DYNAMICS

Unfolding intermediates of the mutant His-107-Tyr of human carbonic anhydrase II†

SRABANI TARAPHDER*, PUSPITA HALDER, TANMOY KUMAR PAUL and SATYAJIT KHATUA

Department of Chemistry, Indian Institute of Technology, Kharagpur, West Bengal 721 302, India
E-mail: srabani@chem.iitkgp.ernet.in

MS received 7 February 2017; revised 31 March 2017; accepted 17 April 2017

Abstract. The mutant His-107-Tyr of human carbonic anhydrase II (HCA II) is highly unstable and has long been linked to a misfolding disease known as carbonic anhydrase deficiency syndrome (CADS). High temperature unfolding trajectories of the mutant are obtained from classical molecular dynamics simulations and analyzed in a multi-dimensional property space. When projected along a reaction coordinate these trajectories yield four distinguishable sets of structures that map qualitatively to folding intermediates of this mutant postulated earlier from experiments. We present in this article a detailed analysis of representative structures and proton transfer activity of these intermediates. It is also suggested that under suitable experimental conditions, these intermediates may be distinguished using circular dichroism (CD) spectroscopy.

Keywords. Carbonic anhydrase; unfolding; marble brain disease; mutant; folding intermediates.

1. Introduction

The three dimensional, native folded structure of a protein is widely known to govern its stability and highly specific biological function.[1–5] Starting with alanine dipeptide[6] as the prototype, remarkable advances have been achieved in the understanding of folding of small proteins.[7] However, for a large protein, the multi-dimensional folding free energy surface is far more complex. Thus, even a rudimentary identification of major folding/unfolding intermediate(s) becomes a computational challenge. In this article, we discuss a computational method for the identification of unfolding intermediates of a large protein.

Carbonic anhydrase (CA),[8] our system of interest, is a ubiquitous enzyme that facilitates the transport and elimination of CO_2 from tissues in all animals, photosynthesizing organisms and in some non-photosynthetic bacteria. The isozyme II of human CA (HCA II) is a moderately large enzyme comprised of 259 amino acid residues. It catalyzes the reversible hydration of CO_2 into bicarbonate[8] with an exceptionally high turnover rate of about 10^6 sec^{-1}. Its native fold is characterized by a highly stable ten-stranded twisted β-sheet (β A to J) at the center flanked by seven α-helices (α A to G). It has a very high thermal stability under physiological condition as evident from its melting temperature, $T_m \sim 58\,^{\circ}C$. It is estimated to become catalytically inactive at ∼60 °C. Figure 1 highlights the residue 107 in the wild type enzyme that is present in a region near but outside the active site. His-107 is a part of hydrogen bonded network that starts from His-119 at the active site and ends at Trp-209 via Glu-117, Tyr-194, Ser-29 and Ser-197.[9] This network contributes to protein stability by holding the large aromatic moiety of Trp-209 residue on the β-core with the N-terminal invariant residue Ser-29.

Replacement of His by Tyr at the residue 107 leads to the loss of at least two hydrogen bonds between the side chains of Glu-117 and His-107. This results in a remarkably destabilizing folding mutation and loss of function that has long been linked to the misfolding disease carbonic anhydrase deficiency syndrome (CADS). The mutant His-107-Tyr of HCA II shows 64% of CO_2 hydration activity compared to that of the wild-type protein at low temperature where the mutant is stable.[10] On the other hand, its stability towards thermal and guanidine hydrochloride (GuHCl) denaturation is found to be highly compromised. Researchers have utilized activity assays, CD, fluorescence, NMR, cross-linking, aggregation measurements and molecular modeling to map the properties of this remarkable mutant.[10] Loss of enzymatic activity occurs at ∼16 °C while its melting temperature, T_m is reduced to ∼22 °C.[10] GuHCl-denaturation

*For correspondence

†Dedicated to the memory of the late Professor Charusita Chakravarty.

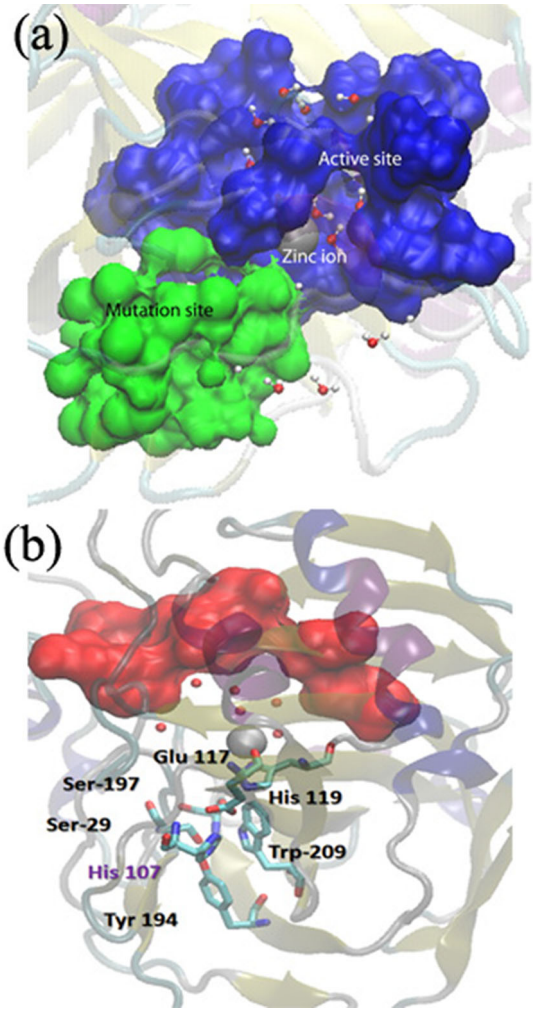

Figure 1. High resolution crystallographic structure of wild type HCA II (PDB id: 2CBA[8]) highliting (**a**) the active site (*blue* surface) with zinc ion (*grey* sphere) and the region around the mutation site of His-107 (*green* surface); (**b**) important amino acid residues interacting with His-107 at the mutation site. Upper portion of the active site region containing His-64 is shown as a red surface in (**b**).

study at 4 °C showed that the native state of the mutant is destabilized by 9.2 kcal mol^{-1}.[19] This is reportedly the greatest degree of protein destabilization measured for a naturally occurring human mutation.[10] However, in the absence of structural data, the correlation between its instability and impaired catalytic function is not clearly understood.

Extensive folding studies on wild type HCA II and its mutants at the residue 107 have led to the postulation of following major intermediates.

(i) For the wild type HCA II, the transition between the folded native (N) and unfolded (U) states is mediated primarily by a partially unfolded intermediate, termed as the molten globule (MG). The MG intermediate has a largely native-like β C-G segment, but no native folded active site. It is catalytically inactive and highly prone to aggregation.

(ii) For the mutant His-107-Tyr of HCA II, an early unfolding intermediate, termed as the molten globule light (MGL) state, is detected preceding the MG intermediate during chemical denaturation at 4 °C. Like MG, it is catalytically inactive, but appears to be less prone to aggregation.

(iii) For Cyp18-assisted refolding of denatured HCA II and its two destabilized mutants (His-107-Asn and His-107-Phe), two kinetic intermediates (MGL_k and MG_k) are detected similar to MGL and MG as mentioned above. A second pathway is involved in the formation of an equilibrium intermediate, MGL_e that is misfolded with reduced catalytic activity. In addition, MGL_e does not exhibit any notable aggregation propensity. However, intermediates equivalent to MGL_k or MGL_e have not been investigated so far in the most unstable mutational variant His-107-Tyr.

Near-UV CD spectroscopic studies have been carried out on MGL/ MG intermediates of the mutant His-107-Tyr of HCA II as a highly sensitive diagnostic tool for the tertiary structure. A model of the wild type structure exhibits strong spectral bands in the near-UV region that remain nearly unchanged if the temperature is raised from 8 to 50 °C.[19] For the mutant, the near-UV CD spectral bands contain all features characteristic of the native conformation of HCA II. However, the mean residue ellipticity, θ is reduced to ∼30% of that of the wild type model. Upon raising the temperature, the magnitudes of the spectral bands decrease. But they do not vanish completely even at 50 °C indicating retention of some residual tertiary structure in the vicinity of some Trp residues even at this elevated temperature. However, no quantitative information on the secondary and tertiary structure contents is available from these studies.

In view of the above, one may put forward a simple hypothesis. Under physiological condition, the mutant His-107-Tyr may be present in a misfolded and catalytically impaired state such as MGL_e. Alternatively, it may be found as a mixture of partially unfolded states closely mimicking the intermediates MGL_e, MGL_k and MG_k. Therefore, determination of structural analogues of these intermediates, if any, in His-107-Tyr appears to be the first step in establishing our hypothesis.

HCA II is a moderately large protein. Identification of potential (un)folding intermediates for this system is thus a non-trivial computational challenge. Unlike small proteins, a single collective system variable, often referred to as the reaction coordinate for folding,

may not correctly map out the (un)folding pathways or the major intermediates along the path. Moreover, our objective is to identify (un)folding intermediates having distinct aggregation propensity and catalytic activity. Following the original proposal by Daggett and Levitt,[11] we have recently used high temperature classical molecular dynamics simulations to identify potential unfolding intermediates of His-107-Tyr of HCA II. Such simulations are well-known to accelerate sampling of different regions of the folding landscape without affecting the pathways.[20] However, relative flatness of the free energy landscape at high temperatures makes it difficult to extract structures of desired properties that would eventually correspond to a folding intermediate. Therefore, a computational methodology has been designed and described in Ref.[12] to address this issue specifically. Different steps adopted in this method are schematically shown in Figure 2 and their salient features are summarized below.

(i) Full atomistic trajectories at high temperatures are used to create a N_p-dimensional property space. This space is constructed by computing along the trajectories N_p properties such as total C_α root-mean-square deviations (RMSD), radius of gyration (R_g), overall solvent accessible surface area (SASA), fraction of native backbone (f_{bb}) and side chain (f_{sc}) contacts, selected backbone dihedral angles, native state-like integrity of the active site and aggregation propensity of several hot spots for aggregation.[12]

(ii) A one-dimensional reaction coordinate, d_{mean} is defined in this property space such that wild type-like structures appear at low values of d_{mean} and those with greater degrees of unfolding are located at progressively higher values of d_{mean}.

(iii) All N_p properties are then simultaneously used as clustering parameters to carry out K-means clustering of different protein conformations along each of the high temperature trajectory. Two major clusters are obtained at different stages of unfolding. The structures corresponding to each of these clusters show up as two overlapping peaks in their respective population histograms projected along d_{mean}. Accordingly, the presence of at least two unfolding intermediates (labelled as I.1 and I.2 and schematically shown in Figure 2) are postulated along the simulated high temperature trajectories.[12]

(iv) Several structures belonging to each of these clusters are further annealed to lower temperatures, equilibrated and their population histogram projected along d_{mean}. It may be expected that the minima in the free energy landscape would become deeper at lower temperatures. Therefore, structures associated with these minima will appear as peaks as high values of d_{mean} when projected along d_{mean}. Therefore, putative structures are chosen by selecting structures corresponding to a narrow range of values of d_{mean} under each peak maximum. This procedure ensures that the population histogram will show better separated peaks along d_{mean} corresponding to persistent minima (albeit local minima) in the free energy landscape.

(v) Four major peaks are identified in the combined population histogram derived from all the trajectories. The structures corresponding to each of these peaks are analyzed and qualitatively mapped on to MGL_e, MGL_k and MG_k.

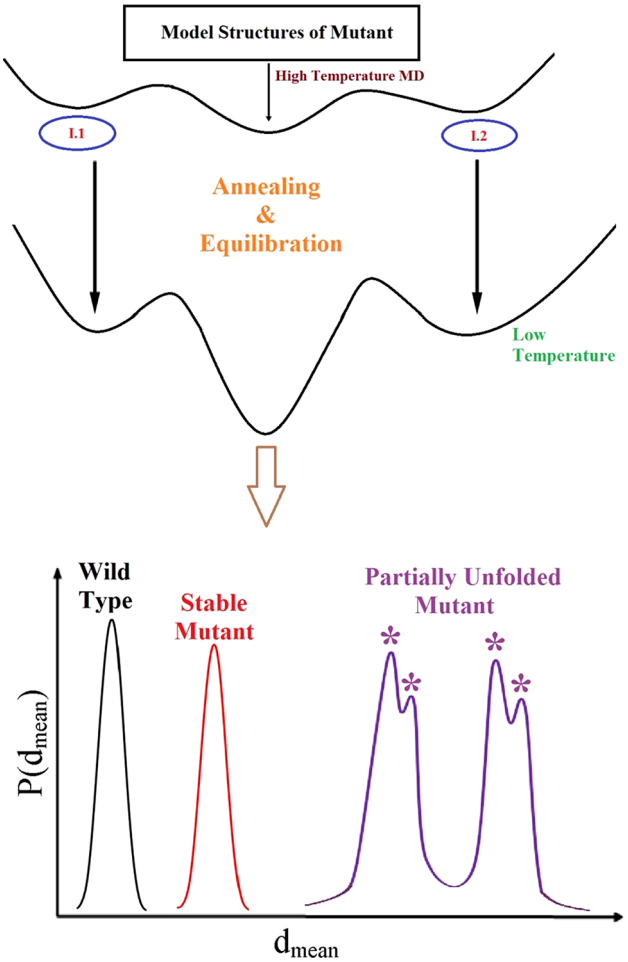

Figure 2. Schematic representation of the computational methodology designed in Ref.[12] First putative unfolding intermediates I.1 and I.2 are detected from high temperature simulations followed by conformational clustering in property space. Several structures are harvested at lower temperatures by annealing those of I.1 and I.2 and projected along the reaction coordinate d_{mean}. Partially unfolded structures that are also potential candidates of unfolding intermediates are associated with the peak maxima at high values of d_{mean}.

Following the above method, four distinct sets of structures, A, B, C1 and C2 are detected as putative unfolding intermediates of the mutant His-107-Tyr of HCA II.[12] The structures belonging to set A are least unfolded and least prone to aggregation. They are found to have a native-like active site to some extent. The structures in sets B, C1 and C2 exhibit progressively higher degrees of unfolding and loss of native active site structure along with increasingly higher propensity to aggregate. Our earlier study[12] was restricted to structure based analysis only. Experiments indicate different catalytic activity of the postulated intermediates. However, at the high temperatures employed in our simulation, protein hydration undergoes very large changes. Therefore, any projection of the catalytic activity based upon the high temperature structures is not expected to be accurate.

In this article, the primary aim is to map the simulated unfolding intermediates to the experimentally postulated ones. For this purpose, selected minimum energy structures are harvested from the sets A, B, C1 and C2 and characterized in terms of several key properties. It is assumed that persistence of a catalytically important proton transfer path at the active site region is a good indicator of the catalytic potential of a detected intermediate structure. Based on this assumption, projected catalytic activities of the selected structures are estimated. We also examine the possibility of detecting the proposed intermediate structures using far and near-UV CD spectroscopy.

The present work, although similar in spirit, does not attempt to perform a replica exchange molecular dynamics simulation. This is because any such project would involve enormous computational cost on account of size of the protein. Even if any such project is taken up, lack of quantitative information folding pathways and structure of intermediates involved may hinder fruitful corroboration to experimental results. Therefore, we focus our search instead on partially unfolded structures that have properties as suggested by experiments.

2. Computational method

Detailed methodology of generating the high temperature unfolding trajectories for the mutant His-107-Tyr has been discussed earlier[12] and is briefly outlined here.

2.1 *Molecular dynamics simulation*

We have utilized in the present study following sets of molecular dynamics trajectories for the wild type and mutant proteins.

(1) Trajectory *wt*: wild type enzyme at 300 K,
(2) Trajectory *mut*: the mutant His-107-Tyr at 277 K where it is expected to be most stable, and
(3) Trajectories *mut-ht*: the mutant His-107-Tyr simulated at 500 K under four different conditions of heating to 500 K.[12]

Each set of trajectory has been generated from an input model structure, solvated, energy minimized and heated to a desired temperature in a large cubic box with around 15,000 TIP3P[15] water molecules with periodic boundary conditions using the solvation utilities of NAMD.[13] In the next step, 1 ns of equilibration and 50 ns of production runs are carried out under isothermal-isobaric conditions using CHARMM22[16] force field parameters. The multiple time step algorithm (RESPA[17]) was used with a MD time step of 1 fs. The particle-mesh-Ewald[18] summation method was applied to treat the electrostatic interactions. The pressure was kept constant at 1 atm using Langevin piston method with a damping coefficient of $5\,ps^{-1}$. The simulation temperature was fixed at the desired value using Langevin damping with a coefficient of $5\,ps^{-1}$. An atom based cutoff of 13.5 Å with switching cutoff at 10 Å was used for non-bonded van der Waals interactions. Upon equilibration, all trajectories exhibit relative energy fluctuations ${}^{\sigma_E}/_{\overline{E}} \sim 10^{-3}$ (data not shown).

2.2 *Structural variation upon partial unfolding*

In order to monitor the variation of secondary structure (SS) elements associated with partial unfolding, the protein structure is divided in to ten different regions as shown in Table 1 and Figure 3. For a given structure, the percentage of α- and 3_{10}-helices within a given region is denoted as $P_H = n_H/n_{H,ref}$, where n_H is the number of amino residues in this region forming the helix compared to that of a reference structure. In the present work, the structure of the wild type protein generated at 300 K at the end of a 50 ns MD run has been used as the reference. Analogous percentage of β-sheets is labelled as P_S. Estimates of P_H and P_S are presented in Table 1 for the wild type protein and its mutant have been estimated from the trajectories *wt* and *mut* using the software STRIDE interfaced with NAMD.[13]

The active site of HCA II is known to be comprised of the amino acid residues 5, 7, 62, 64, 67, 92–96, 106, 117–121 and 143–145. Since the active site is expected to be perturbed during the thermal unfolding process, it is found convenient to define the volume, V_{as} of a partially unfolded structure as that of a sphere that circumscribes all these residues notwithstanding the extent of deviations from the wild type-like active site structure. The center of this sphere is assumed to coincide with the center of mass of C_α-atoms of all the residues listed above. The radius of this circle is approximated by the distance between this center of mass and the C_α-atom located farthest from it.

Table 1. Region-wise partitioning of secondary structure (SS) elements of wild type HCA II (*wt*) and its mutant His-107-Tyr (*mut*).

Region	Residue number	SS (*wt*)	*wt*		*mut*		A		B		C1		C2	
			P_H	P_S	P_H	P_S	P_H	P_S	P_H	P_S	P_H	P_S	P_H	P_S
1	3–20	α A	38.9	0.0	3.7	0.0	38.89	0.0	0.0	0.0	44.45	0.0	44.45	0.0
2	21–30	α B	0.0	10.0	0.0	9.6	70	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	31–125	β B,C,D,E,J β a,b,c,d	0.0	50.5	1.7	46.6	4.21	44.21	4.21	42.10	4.21	53.70	0.0	54.73
4	126–138	α C, D	38.5	7.7	37.5	7.7	46.15	0.0	46.15	0.0	53.84	0.0	30.77	0.0
5	139–150	β F	0.0	75.0	0.0	75.0	0.0	83.34	0.0	83.34	0.0	83.34	0.0	83.34
6	151–170	α E	60.0	10.0	43.2	10.0	60	5	70	5	0.0	5	70	5
7	171–180	β A	0.0	20.0	15.8	23.0	0.0	40	0.0	50	0.0	50	0.0	50
8	181–216	α F, β G,H	0.0	41.7	1.8	43.6	13.89	55.55	0.0	55.56	0.0	52.78	11.12	41.67
9	217–230	α G	57.1	0.0	56.9	0.0	71.43	21.43	64.30	21.43	50	21.43	71.43	21.43
10	231–261	β I	0.0	7.1	0.0	6.6	16.13	19.35	0.0	9.67	0.0	9.67	0.0	9.67

P_H and P_S are the percentages of α-, 3_{10}-helices and β-sheets, respectively, in a given region. Also shown are percentages of SS elements in the structures A, B, C1 and C2 that represent potential unfolding intermediates of the mutant.

2.3 *Construction of multi-dimensional property space and reaction coordinate*

Several properties were monitored along each trajectory listed above and 35 of them were found to exhibit substantial dynamical variation in the trajectories *mut-ht* compared to the two reference trajectories, *wt* and *mut*. Selected few of these, relevant to the present study, are listed in Table 2. A complete list is available in Ref.[12]

The extent of unfolding of a given structure is conventionally monitored in terms of structural parameters such as RMSD, R_g and solvent accessible surface area (SASA). We have used the change in SASA (compared to *wt* or *mut*) as one of the indicators of increasing aggregation propensity displayed by a given structure. In addition, any change in structural integrity of the active site is characterized in terms of (i) tilt angle θ_t and (ii) distance d_{as}. The tilt angle is defined as the angle between two vectors drawn from the C_α-atom of Ser-105, first to the zinc ion and second to the C_α-atom of residue 107. The distance d_{as} refers to the normal distance between zinc ion and the plane containing C_α-atoms of its three coordinated ligands, His-94, His-96 and His-119. If a stable coordination environment is maintained around the zinc ion (as in *wt* or *mut*), θ_t lies between 23° and 48° or 73° and 95° and $3.0 \lesssim d_{as} \lesssim 5.2$ (in Å).[12]

The property space is also used to define a multidimensional-embedded, one-dimensional reaction coordinate, d_{mean} as[12]

$$d_{mean}{}^2(t_i) = \frac{1}{N_p N_t} \sum_{j=1}^{N_p} \sum_{k=1}^{N_t} \left[X_j(t_i) - X_j{}^{ref}(t_k) \right]^2$$

In this notation, N_p properties span the property space. The mean distance, $d_{mean}(t_i)$ corresponds to the distance in N_p-dimensional property space at time t_i between a given dynamical trajectory and a reference trajectory. N_t denotes the number of time slices along each of them. While computing $d_{mean}(t_i)$, the j-th property $X_j(t_i)$ is compared to the entire range of values of the same property, $X_j{}^{ref}$ exhibited by a reference system. We have used a 50 ns long MD trajectory of the wild type enzyme at 300 K as the reference. Therefore, structures representing the stable native fold of HCA II are associated with a narrow distribution near $d_{mean} = 0$. Structures with higher degree of unfolding compared to the native fold are expected to appear at progressively higher values of d_{mean}.

2.4 *Detection of unfolding intermediates*

The high temperature trajectories are subjected to K-means clustering using $N_p = 35$ properties as clustering parameters in the multi-dimensional property space.[12] A potential unfolding intermediate is detected as the cluster with maximum mean silhouette value. As mentioned earlier, these clusters appear at higher values of d_{mean} (compared to the wild type) when projected along the reaction coordinate.

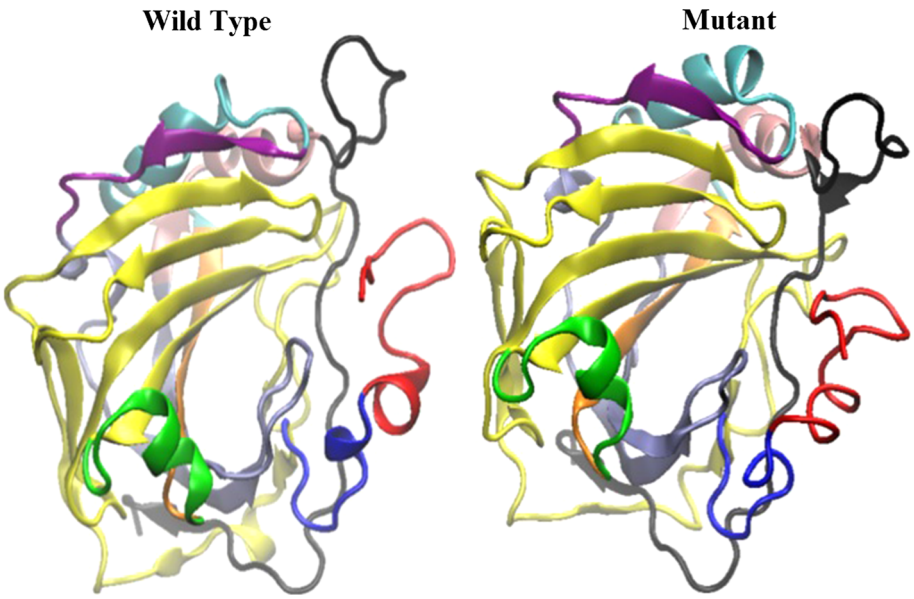

Figure 3. Representation of the structures *wt* and *mut* of the wild type and mutant protein from the terminal snapshots of the respective 50 ns long trajectories. Also shown are the ten different regions chosen for secondary structure segments as 1 (*red*), 2 (*blue*), 3 (*yellow*), 4 (*green*), 5 (*orange*), 6 (*cyan*), 7 (*purple*), 8 (*ice blue*), 9 (*pink*) and 10 (*black*).

Table 2. Selected properties of wild type HCA II (*wt*) and its mutant His-107-Tyr (*mut*) in the multi-dimensional property space.

	wt	*mut*	A	B	C1	C2
1 Total RMSD (Å)	0.0	0.87	2.17	3.17	3.44	5.89
2 Radius of gyration (R_g in Å)	38.5	38.3	42.1	42.3	42.8	42.3
3 Total SASA (Å^2)	57,235.8	54,662.8	123,955.4	126,342.5	141,178.5	125,223.2
4 Tilt angle θ_t (°)	79.2	70.3	90.64	95.28	81.09	87.12
5 d_{as} (Å)	4.36	5.76	4.74	4.30	2.21	2.86
6 Fraction of backbone contacts	–	–	0.18	0.19	0.18	0.17
7 Fraction of side chain contact	–	–	0.80	0.72	0.77	0.67

Also shown are the estimated values of the same for representative structures A, B, C1 and C2 belonging to potential unfolding intermediates of the mutant. The fraction of contact is calculated with respect to the stable mutant structure *mut* with a cutoff distance equal to 6 Å (backbone) and 8.5 Å (side chain).

2.5 *Identification of proton transfer path*

The rate determining catalytic step for HCA II is an intramolecular proton transfer from a zinc-bound water molecule to the side chain of His-64 (located ~10 Å away on the other side of the active site). A highly efficient proton transfer in the wild type enzyme is mediated by a hydrogen bonded network of 2–3 water molecules at the active site. In the present work, we have chosen four minimum energy structures, one each from the sets A, B, C1 and C2. Each structure is then solvated, energy minimized and equilibrated at 300 K using the solvation utilities of NAMD.[13] The protein atoms were kept fixed to obtain a reasonable approximation of the active site hydration without altering the protein structure. For this purpose, we enlist the heavy atoms of all polar amino acid residues and O atoms of water molecules as probable nodes in the proton relay and classify them in hydrogen bonded clusters using an upper cutoff of 3.5 Å.[9] A *catalytically important* proton path is detected when side chain ($N_{\delta 1}$ and $N_{\varepsilon 2}$) atoms of His-64 and O of the zinc-bound water molecule belong to the same cluster. Therefore, the projected catalytic activity of a structure is assumed as high if any analogue of the key proton transfer pathway persists in it.

To have a better solvation profile of these equilibrated structures that were extracted from trajectories at 500 K, additional dynamical trajectories X-*ext-solv* (X = A, B, C1, C2) were allowed to evolve for another for 2–5 × 10^5 steps at 300 K starting from structure X and ***keeping all the protein atoms***

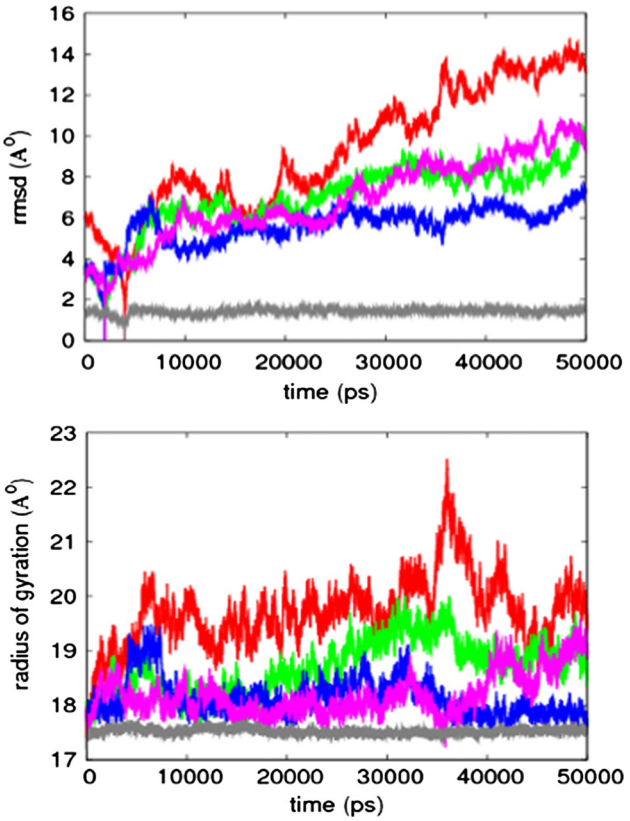

Figure 4. Variation of overall C_α root-mean-square deviation (RMSD) (*top*) and radius of gyration (R_g) (*bottom*) along the 50 ns long trajectories corresponding to the wild type protein at 300 K (*wt*, *grey*) and four different mutant trajectories at 500 K (*mut-ht*; *red*, *green*, *blue* and *pink*).

fixed. These trajectories were then used to generate occupancy plots for the water oxygen atoms created using the *volmap* tool in VMD.[21] These plots are useful in detecting transient hydrogen bonds that may complete the key proton transfer pathways for a short time.

2.6 *Predicted circular dichroism (CD) spectra of intermediate structures*

Theoretical estimation of CD spectra of wild type HCAII and its mutant His-107-Tyr has been carried out by using the matrix method with *ab-initio* monopoles implemented within *DichroCalc*.[14]

3. Results and Discussion

3.1 *Reference structures of wild type and mutant protein*

We have obtained the structures *wt* and *mut* of the wild type and mutant protein from the terminal snapshots of the respective 50 ns long trajectories. These two are used in our analysis as reference for the highly stabilized native fold of wild type HCA II and the most stable structure that could be modeled for the His-107-Tyr mutant. They are presented in Figure 3 where we have also highlighted using different colors the ten regions defined in Table 1. Variation of the protein structure along the trajectories *wt and mut-ht* has been shown in Figure 4 in terms of total C_α root mean square deviation (RMSD) and radius of gyration (R_g).

3.2 *Putative unfolding intermediates*

Quite a few significant transitions are observed in Figure 4 that could correlate to transitions along different folding pathways. However, as discussed at length in Ref.[12], use of one or two such properties as clustering parameter are found to yield overlapping clusters from the high temperature trajectories *mut-ht*. To address this issue, conformational clustering was carried out using 35 properties as clustering parameters to obtain clusters having substantially reduced overlap. Two clusters thus obtained are labeled as I.1 and I.2.

Subsequently, the structures belonging to the sets I.1 and I.2 are annealed fast to lower temperatures (290 and 310 K, below and above the melting temperature of 295 K of the mutant) and equilibrated to harvest a large number of replicas of possible intermediate structures. Structures from different trajectories are combined and projected along d_{mean}. The resultant population histograms are shown in Figure 5 for a few representative sets. For every such structure set compiled, the distribution is scanned for predominant peaks (such as the peaks A-F as shown in Figure 5). A narrow interval of d_{mean} is chosen around the position of each peak and the structures contributing to this interval by different trajectories are collected in separate sets. These structure sets are evidently well separated along d_{mean} and are highly probable in comparison to other structures sampled. They are further subjected to analysis to determine their aggregation propensity and integrity of the active site. This procedure yields four sets of partially unfolded mutant structures with the following properties.[12]

1. **Set A**: 6455 partially unfolded mutant structures are harvested that have both active site and inner β-core intact and a low propensity to aggregate.
2. **Set B**: 8320 partially unfolded structures that have lost the active site integrity but still retain a nearly invariant inner β-core. They are also characterized by

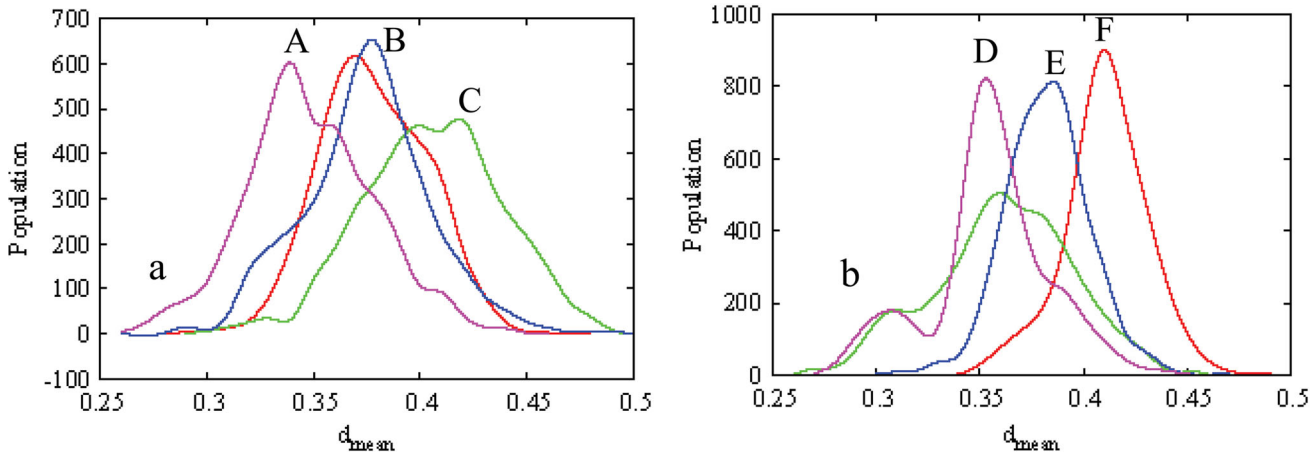

Figure 5. Population histogram of representative structure sets along d_{mean} considered as potential candidates for unfolding intermediates of the mutant. Different *colors* are used to indicate that the underlying structures have been harvested by annealing different sets of structures belonging to I.1 or I.2.

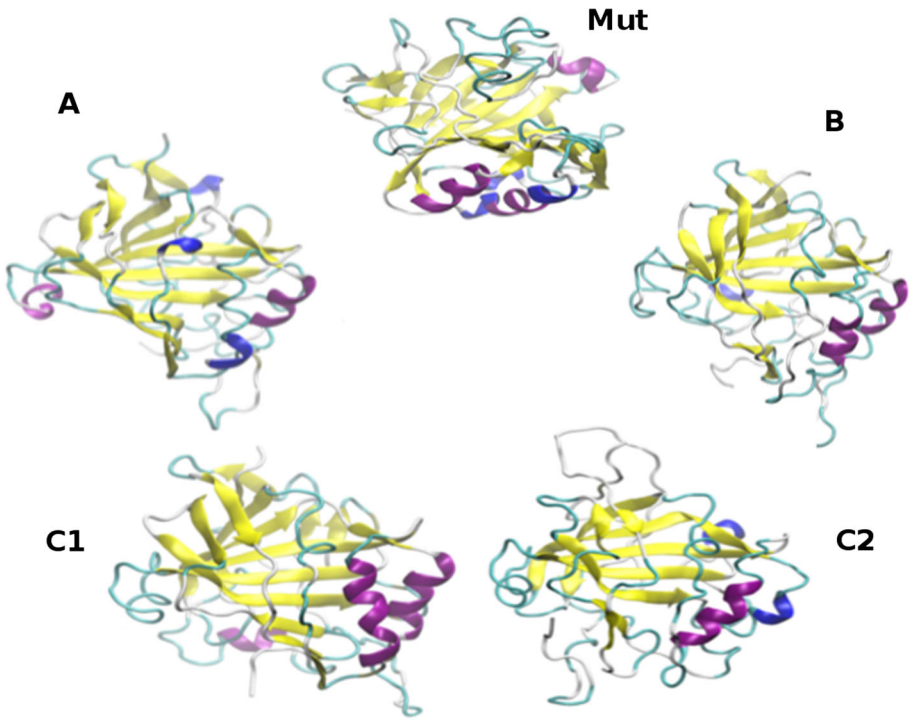

Figure 6. Most stable structure of the mutant derived at 277 K (*Mut*) and partially unfolded structures A, B, C1 and C2 representing putative unfolding intermediates of the mutant.

Table 3. Calculated active site volume for the wild type protein (structure *wt*) and its mutant His-107-Tyr (structure *mut*) compared to the partially unfolded mutant structures A, B, C1 and C2.

Structure	*wt*	*mut*	A	B	C1	C2
Active site volume, V_{as} (Å^3)	12,502.92	12,628.49	14,435.84	31,464.82	7188.61	12,065.91

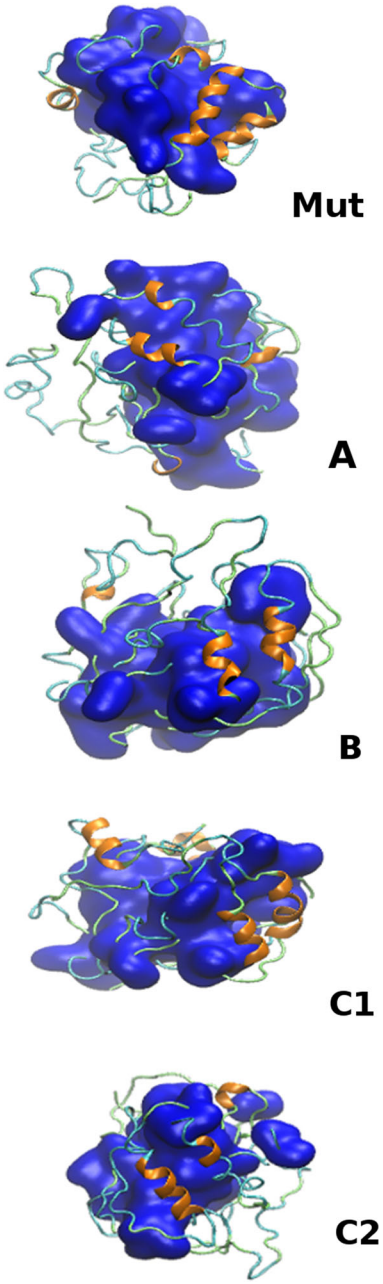

Figure 7. Variation of peripheral helices (*orange*) around the hydrophobic β-core (*blue*) in the structures *mut*, A, B, C1 and C2.

relatively higher aggregation propensity compared to those of Set A.

3. **Set C1 and C2**: These structures are substantially unfolded compared to the sets A and B, do not satisfy the conditions of the active site integrity, show the onset of unfolding of the inner β-core and exhibit high aggregation propensity. The 6989 structures belonging to set C1 exhibit on an average smaller deviation from the conditions of stable active site structures in comparison to the 7718 structures belonging to set C2.

Evidently, sets A and B appear to corroborate to the properties of proposed intermediates MGL_e and MGL_k qualitatively, while sets C1 and C2 appear to mimic MG intermediate. Average properties calculated for each of these sets have been presented earlier [12] and they justify such correlation.

3.3 *Analysis of representative mutant structures*

We have next selected the minimum energy structure from each of the four sets used in our subsequent analysis. These four representative structures are labeled as A, B, C1 and C2 from the corresponding sets of partially unfolded structures of the mutant. These are presented in Figure 6.

It is clear from Figure 6 that unfolding of *mut* leads to variation in the secondary structure elements. The percentages, P_H and P_s of helical and β-sheet content, respectively, in each of the ten regions are presented in Table 1 and compared with those obtained for *wt* and *mut*. As expected for the regions near the N-terminus, regions 1 and 2 exhibit large variations with the α A and B helices that are lost in structure B, but eventually reforms in C1 and C2. α C and D helices are not seen beyond the structure B. α E helix persists in structures A, B and C2 while being completely unfolded in structure C1. Variation of peripheral helices around the β-core are highlighted for all four structures in Figure 7.

Stability of the β-core is evident from the little or no variation in the regions 3 (β B, C, D, E, J), 5 (β F) and 8 (β G, H). The onset of perturbation of the stable hydrophobic core is indicated mainly by the regions 7 and 9 (associated with the loss of β A and α G) along with small fluctuations in region 3. The associated changes in active site are well-reflected in the calculated values of active site volume, V_{as} presented in Table 3. Space-filling models of the residues supposed to be forming the active site are presented in Figure 8 for the mutant structures A, B, C1 and C2 as a visual guide. The structures *wt* and *mut* have active site volumes within about 1% of each other. Structures A and C2 exhibit only marginal deviations in V_{as} from the reference structures. Remarkably, V_{as} of structure B becomes nearly 1.5 times that of *wt*. The active site appears to shrink considerably in C1. It is beyond the scope of present work (limited to only four structures) to establish if this is an artifact of our simulations or whether these observations can be statistically significant. At this stage, no apparent correlation

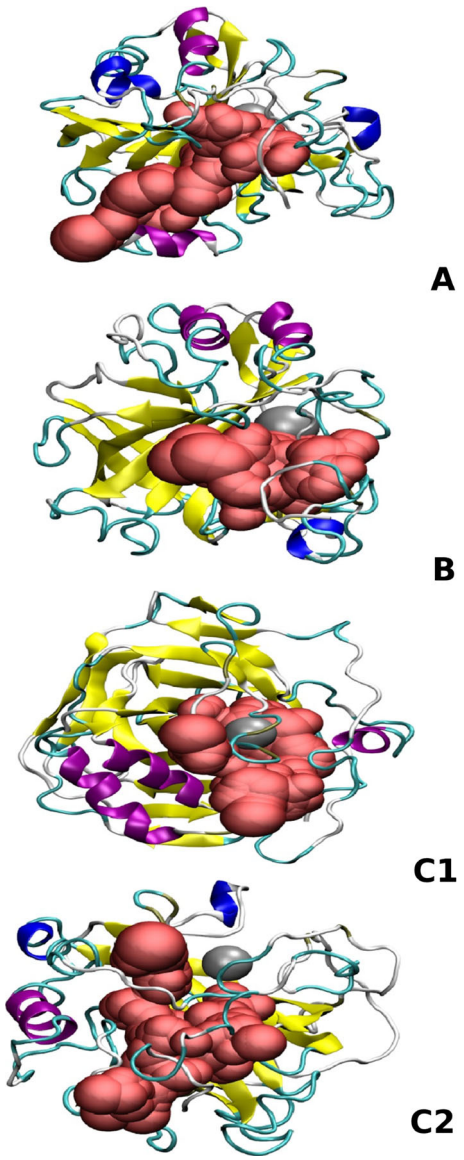

Figure 8. Van der Waals surfaces of all amino acid residues considered for the calculation of active site volume, V_{as} for the partially unfolded mutant structures A, B, C1 and C2.

is evident between V_{as} and retention of wild type-like environment at the active site.

Table 2 summarizes some key properties of A, B, C1 and C2. The variation of RMSD, R_g and total SASA values evidently indicates an increasing degree of unfolding as we move from structure A to C1 and C2. The increase in total SASA is found to be maximum in the structure C1. With increasing unfolding, regions 3 and 9 are found to undergo most prominent changes in their SASA values (∼5%) in the structure C2 compared to the structure *mut*. These data are consistent with the observed values of fraction of native backbone and side chain contacts in these structures.

It has been shown in Table 2 that the structural features of set A resemble that of the mutant most closely. Deviations of the active site from its most stable configuration are more prominent in the structures C1 and C2. In particular, the active site shows signs of collapsing as evident from rather small value of d_{as} for set C2. In addition, only about 18% hydrophobic C_α–atom contacts are retained in these structures compared to those present in the most stable mutant structure, *mut*.

3.4 *Hydrogen bonded network at the active site*

The solvation environment of the catalytic zinc ion changes as the zinc-bound water molecule forms a small hydrogen bonded cluster with Glu-106 and Glu-117 along with His-96 (structures B and C1) and with His-119 (structure C2). His-64, the key catalytic residue, points away from the active site in the structures B, C1 and C2. This, coupled to the depletion of active site water molecules, makes formation of stable proton transfer paths highly unlike in these structures. Therefore, the structures B, C1 and C2 appear to mimic the loss of catalytic activity and increased propensity of aggregation exhibited by the intermediates MGL_k and MG_k. The residue His-64 points to the active site cavity in structure A, thus retaining its optimal side chain orientation for an efficient transfer of proton. However, structure A is found to lack water molecules that can reside at the active site long enough to form the proton transfer path between the zinc bound water and His-64.

It is natural to question the validity of the analysis presented above given the fact that these structures were generated from high temperature MD trajectories and therefore, would not represent the hydration structure around the active site accurately. Therefore, we further analyzed the hydration of this region at 300 K along extended trajectories generated from the structures. When equilibrated at 300 K for 3×10^5 steps more keeping the protein atoms fixed, several water molecules are found to populate the active site. The isosurface corresponding to 50% occupation probability of water molecules within a distance of 8 Å from the catalytic zinc ion has been shown in Figure 9. Extended regions of hydration are observed that may lead to the formation of proton paths transiently at a much longer time. Proton transfer under such circumstances is expected to encounter a higher free energy barrier in A compared to *wt* or *mut*. This corroborates well with a lower catalytic activity of this class of structures and hence the latter seem to agree with the features of intermediate MGL_e (Figure 10).

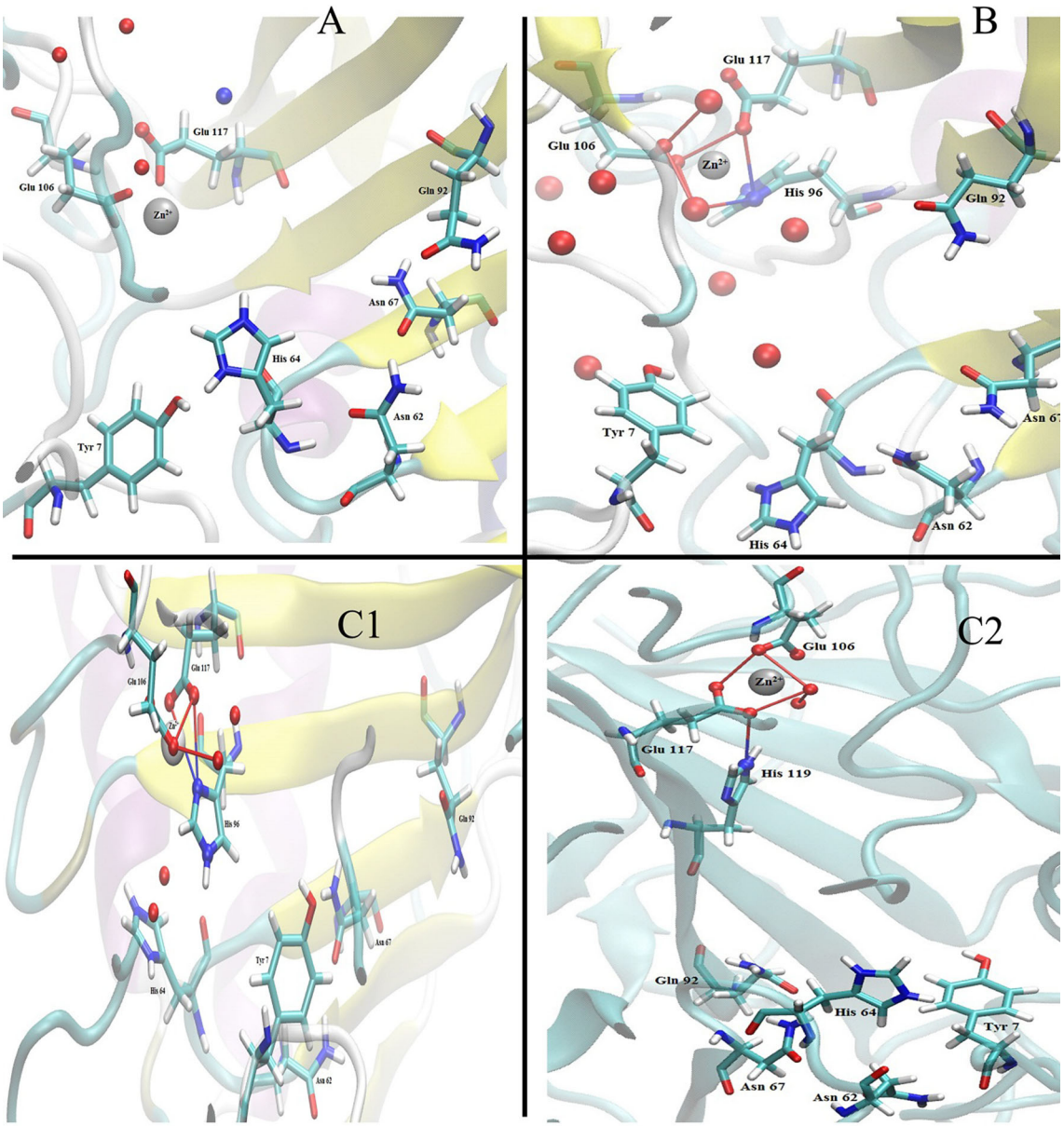

Figure 9. Lack of hydrogen bonded proton path between the zinc-bound water and His-64 at the active site of the unfolding intermediates A, B, C1 and C2 of the mutant His-107-Tyr of HCA II. Catalytically important active site residues are shown using sticks and O-atoms of active site water molecules are shown as red spheres. Hydrogen bonds between network nodes are highlighted using *red* sticks.

The change in structure around the active site of the intermediates C1 and C2 is fairly evident from the observed deviation (Table 3) of the associated tilt angle θ_t and distance d_{as} compared to those of the structures *wt* and *mut*. Alignment of the structures C1 and C2 with respect to the zinc ion indicates a significant perturbation of the active site region especially for residues Trp-5, Tyr-7 and His-64. This observation correlates well with the calculated values V_{as} of structures C1 and C1 as shown in Table 3.

We also investigated the hydration structure by allowing the high temperature structures C1 and C2 to equilibrate at 300 K for 2–5 $\times$ 10^5 steps ***keeping all the protein atoms fixed***. The results are shown in Figure 11 for the structures C1-*ext-solv* and C2-*ext-solv*, respectively. A partial collapse of the tetrahedral coordination environment of the zinc ion is more prominent in C2 than in C1. This results in the expulsion of water molecules from the erstwhile active site. The isosurfaces corresponding to 50% occupation probability by water molecules are found to be concentrated around the zinc ion along with Glu-106 and Glu-117 in both the cases. It therefore appears that both these structures are incapable of showing any catalytic activity.

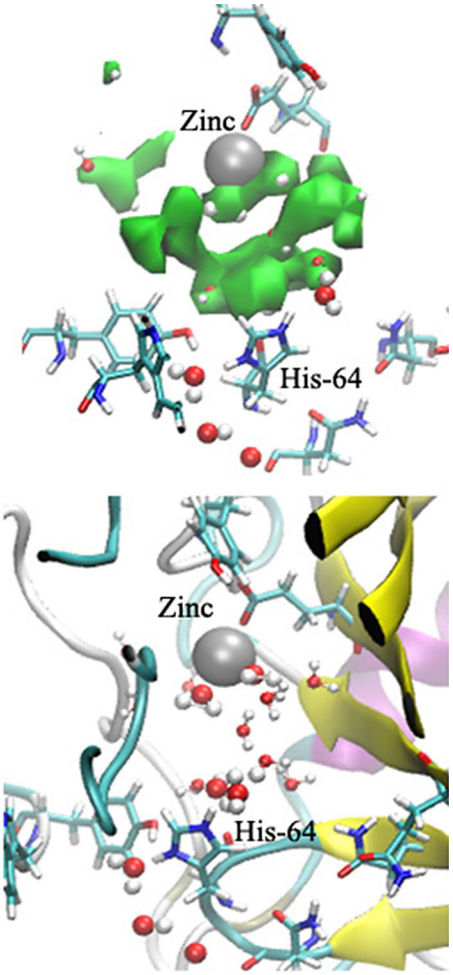

Figure 10. Hydration of the active site of the unfolding intermediate (structure A) of the mutant His-107-Tyr of HCA II showing the isosurface with 50% occupation probability by water molecules (*upper panel*) and a molecular representation of the active site hydration at long times keeping the protein structure fixed (*lower panel*).

3.5 *Calculated CD spectra*

We have presented in Figure 12 theoretically predicted CD spectra of the structures *wt*, *mut*, A, B, C1 and C2 at both far- and near-UV regions. Structure A gives the most distinct spectral band in the near-UV region compared to either wild type or the most stable mutant structure. As pointed out earlier, the observed changes with higher degree of unfolding seem to indicate the underlying changes in tertiary contacts around the Trp residues present in these structures. However, the results presented in this figure take into consideration only one structure from each set of intermediates. Therefore, the trends shown here are merely indicative of the fact if the mutant can eventually be experimentally trapped, these data may be used to estimate relative populations of MGL_e, MGL_k or MG-like structures in it.

4. Conclusions

In this article, we have utilized a recently proposed computational methodology to detect and characterize potential unfolding intermediates of a moderately large protein His-107-Tyr mutant of HCA II. It is found that a misfolded intermediate with lowest degree of unfolding and relatively lower SASA is expected to exhibit catalytic activity. However, such activity will depend on the dynamics of hydration of the active site. The other intermediates show marked deviation from the both wild type and stable mutant structures. Unfolding seems to remove water molecules from the active site, thereby rendering them catalytically inactive. It is also suggested that near-UV CD spectra will be helpful in distinguishing these structures.

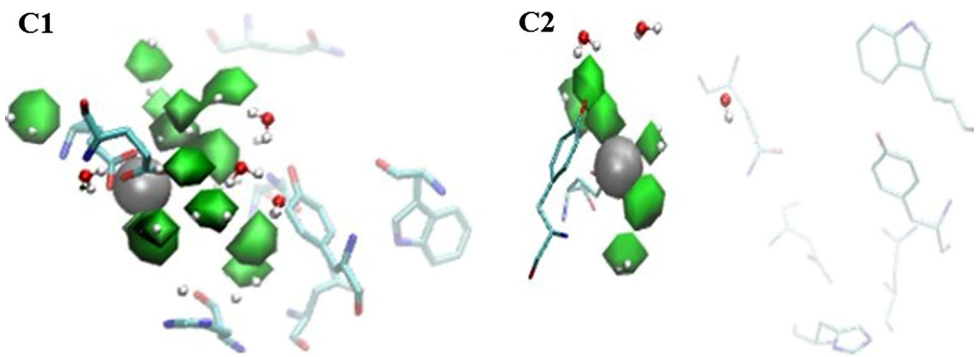

Figure 11. Comparison of isosurfaces (*green*) corresponding to 50% probability of occupation by water molecules from prolonged equilibration of the structure C1 and C2 showing lack of water molecules in the erstwhile active site.

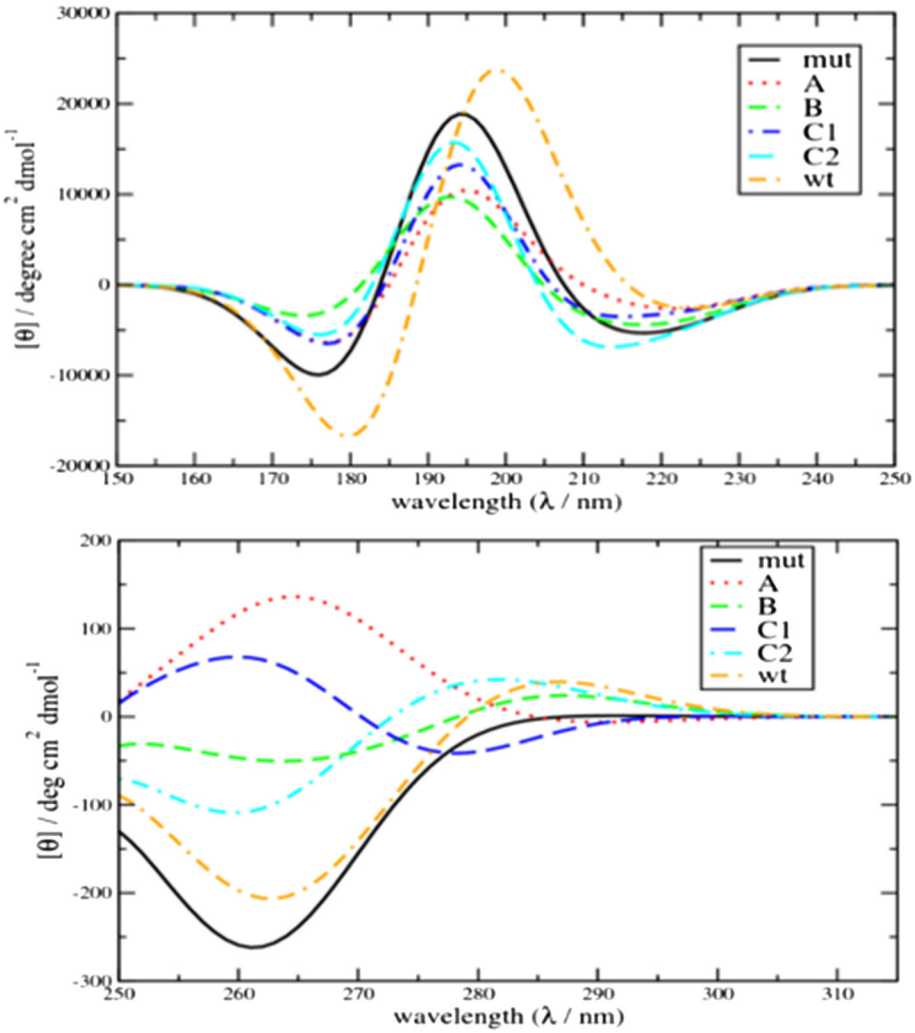

Figure 12. Far-UV (*top*) and near-UV (*bottom*) CD spectra predicted for wild type HCA II (*wt*) and its mutant (*mut*). Also shown are the calculated spectral bands of potential unfolding intermediates A, B, C1 and C2 of the mutant.

The work presented here has been restricted to selected structures for each class of intermediates detected in the multi-dimensional property space. Our results seem to indicate that further insight into the mechanism of unfolding of the mutant and its correlation to its reduced catalytic activity may be probed if optimum pathways between these intermediate states may be developed. This will be addressed in near future.

Acknowledgements

A part of this work has been funded by Council of Scientific and Industrial Research (CSIR), India, Grant number: 01(2485)/11/EMR-II. All computational studies were carried out using the high performance computing facility at Department of Chemistry, IIT Kharagpur, funded by DST-FIST (SR/FST/CSII-011/2005), India. Research fellowships from UGC, India (TKP) and IIT Kharagpur (SK) are gratefully acknowledged.

References

1. (a) Wolynes P G 2015 Evolution energy landscapes and the paradoxes of protein folding *Biochimie* **119** 218; (b) Onuchic J N and Wolynes P G 2004 Theory of protein folding *Curr. Opin. Struc. Biol.* **14** 70
2. (a) Ingolfsson H I, Lopez C A, Uusitalo J J, De Jong D H, Gopal S M, Periole X and Marrink S J 2014 The power of coarse graining in biomolecular simulations *WIRES-Comp. Mol. Sci.* **4** 225; (b) Best R B, Hummer G and Eaton W A 2013 Native contacts determine protein folding mechanisms in atomistic simulations *Proc. Natl. Acad. Sci. USA* **110** 17874; (c) Bowman G R and Pande V S Protein folded states are kinetic hubs 2010 *Proc. Natl. Acad. Sci. USA* **107** 10890
3. (a) Woodside M T and Block S M 2014 Reconstructing folding energy landscapes by single-molecule force spectroscopy *Ann. Rev. Biophys*. **43** 19; (b) Dasgupta A, Udgaonkar J B and Das P 2014 Multistage unfolding of an SH3 domain: An initial urea-filled dry molten globule precedes a wet molten globule with non-native structure *J. Phys. Chem. B* **118** 6380; (c) Chung H S and Eaton W

A 2013 Single-molecule fluorescence probes dynamics of barrier crossing *Nature* **502** 685

4. (a) Roy S and Bagchi B 2014 Comparative study of protein unfolding in aqueous urea and dimethyl sulfoxide solutions: Surface polarity, solvent specificity, and sequence of secondary structure melting *J. Phys. Chem. B* **118** 5691; (b) Ghosh R, Roy S and Bagchi B 2014 Multidimensional free energy surface of unfolding of HP-36: Microscopic origin of ruggedness *J. Chem. Phys.* **141** 135101; (c) Mandal M and Mukhopadhyay C 2014 Microsecond molecular dynamics simulation of guanidinium chloride induced unfolding of ubiquitin *Phys. Chem. Chem. Phys.* **16** 21706; (d) Bhattacharjee N, Rani P and Biswas P 2013 Capturing molten globule state of alpha-lactalbumin through constant pH molecular dynamics simulations *J. Chem. Phys.* **138** 095101
5. (a) Gupta M, Chakravarty C and Bandyopadhyay S 2016 Sensitivity of protein glass transition to the choice of water model *J. Chem. Theory Comput.* **12** 5643; (b) Gupta M, Nayar D, Chakravarty C and Bandyopadhyay S 2016 Comparison of hydration behavior and conformational preferences of the Trp-cage mini-protein in different rigid-body water models *Phys. Chem. Chem. Phys.* **18** 32796; (c) Nayar D and Chakravarty C 2014 Sensitivity of local hydration behaviour and conformational preferences of peptides to choice of water model *Phys. Chem. Chem. Phys.* **16** 10199
6. Brooks C L and Case D A 1993 Simulations of peptide conformational dynamics and thermodynamics *Chem. Rev*. **93** 2487
7. (a) Onuchic J N, Nymeyer H, Garcia A E, Chahine J and Socci N D 2000 The energy landscape theory of protein folding: Insights into folding mechanisms and scenarios *Adv. Protein Chem.* **53** 87; (b) Dill K A, Bromberg S, Yue K, Fiebig K M, Yee D P, Thomas P D and Chan H S 1995 Principles of protein folding – a perspective from simple exact models *Protein Sci.* **4** 561; (c) Duan Y and Kollman P A 1988 Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution *Science* **282** 740
8. (a) Lindskog S 1997 Structure and mechanism of carbonic anhydrase *Pharmacol. Ther.* **74** 1; (b) Christianson D W and Fierke C A 1996 Carbonic anhydrase: Evolution of the zinc binding site by nature and by design *Acc. Chem. Res.* **29** 331; (c) Silverman D N and Lindskog S 1988 The catalytic mechanism of carbonic anhydrase—implications of a rate-limiting protolysis of water *Acc. Chem. Res.* **21** 30; (d) Hakansson K, Carlsson M, Svensson L A and Liljas A 1992 Structure of native and apo carbonic anhydrase II and structure of some of its anion-ligand complexes *J. Mol. Biol.* **227** 1192
9. (a) Roy A and Taraphder S 2008 A theoretical study on the detection of proton transfer pathways in some mutants of human carbonic anhydrase II *J. Phys. Chem. B* **112** 13597; (b) Roy A and Taraphder S 2007 Identification of proton-transfer pathways in human carbonic anhydrase II *J. Phys. Chem.* **111** 10563
10. Almstedt K, Lundqvist M, Carlsson J, Karlsson M, Persson B, Jonsson B-H, Carlsson U and Hammarström P 2004 Unfolding a folding disease: Folding, misfolding and aggregation of the marble brain syndrome-associated mutant H107Y of human carbonic anhydrase II *J. Mol. Bio*. **342** 619
11. Daggett V and Levitt M 1992 Molecular dynamics simulations of helix denaturation *J. Mol. Biol.* **223** 1121
12. Halder P and Taraphder S 2016 Identification of putative unfolding intermediates of the mutant His-107-tyr of human carbonic anhydrase II in a multidimensional property space *Proteins Struct. Funct. Bioinf.* **84** 726
13. Phillips J C, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel R D, Kale L and Schulten K 2005 Scalable molecular dynamics with NAMD *J. Comput. Chem.* **26** 1781
14. (a) Oakley M T, Bulheller B M and Hirst J D 2006 First-principles calculations of protein circular dichroism in the far-ultraviolet and beyond *Chirality* **18** 340; (b) Bulheller B M and Hirst J D 2009 The DichroCalc web interface for dichroism calculations *Bioinformatics* **25** 539; (c) Besley N A and Hirst J D 1999 Theoretical studies toward quantitative protein circular dichroism calculations *J. Am. Chem. Soc.* **121** 9636
15. Jorgensen W L, Chandrasekhar J, Madura J D, Impey R W and Klein M L 1983 Comparison of simple potential functions for simulating liquid water *J. Chem. Phys.* **79** 926
16. (a) MacKerell A D, Bashford D, Bellott M, Dunbrack R L, Evanseck J D, Field M J, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau F T K, Mattos C, Michnick S, Ngo T, Nguyen D T, Prodhom B, Reiher W E, Roux B, Schlenkrich M, Smith J C, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin and Karplus M 1998 All-atom empirical potential for molecular modeling and dynamics studies of proteins *J. Phys. Chem. B* **102** 3586; (b) MacKerell A D, Feig M and Brooks III C L 2004 Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations *J. Comp. Chem.* **25** 1400
17. Tuckerman M E, Berne B J and Rossi A 1990 Molecular dynamics algorithm for multiple time scales: Systems with disparate masses *J. Chem. Phys.* **94** 1465
18. Di Pierro M, Elber R and Leimkuhle B 2015 A stochastic algorithm for the isobaric-isothermal ensemble with ewald summations for all long range forces *J. Chem. Theory Comput*. **11** 5624
19. Almstedt K, Lundqvist M, Carlsson J, Karlsson M, Persson B, Jonsson B H, Carlsson U and Hammarström P 2004 Unfolding a folding disease: Folding, misfolding and aggregation of the marble brain syndrome-associated mutant H107Y of human carbonic anhydrase II *J. Mol. Biol*. **342** 619
20. Jong D D, Riley R, Alonso D O V and Daggett V 2002 Probing the energy landscape of protein folding/unfolding transition states *J. Mol. Biol*. **319** 229
21. Humphrey W, Dalke A and Schulten K 1996 VMD: Visual molecular dynamics *J. Mol. Graphics* **14** 33