Supporting Information:
Structure and Dynamics of the Isozymes II and IX of Human Carbonic Anhydrase

Divya Rai, Satyajit Khatua, and Srabani Taraphder*

Department of Chemistry, Indian Institute of Technology Kharagpur 721302, India

E-mail: srabani@chem.iitkgp.ac.in
Supplementary Information

Determination of starting structure of the catalytic domain of HCA IX

Among the existing high-resolution crystallographic structures deposited in the protein data bank, the structures closest to the catalytic domain of the wild-type HCA IX are the PDB identities: 3IAI$^{S1}$ (1 mutation), 4ZAO$^{S2}$ (7 mutations), 5DVX$^{S3}$ (6 mutations) and 6FE2$^{S4}$ (1 mutation). The PDB code 6FE2 was chosen to be the starting structure. Utilizing the Multiseq$^{S5}$ analysis of Visual molecular dynamics (VMD)$^{S6}$ package, and sequential superimposition of the structures have been carried out, keeping 6FE2 as the reference highlighted in Table S1 and Figure S1. The $Q_H$ value is a metric for structural homology. The $Q_H$ value measures structural conservations where $Q_H=1$ implies that structures are identical and $0.1<Q_H<0.3$ implies poorly aligned structures.

Table S1: RMSD, percentage of sequence and structural similarity and $Q_H$ value calculated for the superimposition of the high resolution crystal structure of the catalytic domain of HCA IX with PDB code 6FE2$^{S4}$ with 3IAI,$^{S1}$ 4ZAO,$^{S2}$ and 5DVX.$^{S3}$

|          | 3IAI  | 4ZAO  | 5DVX  |
|----------|-------|-------|-------|
| RMSD (Å) | 0.4308| 1.675 | 0.6608|
| Percentage identity | 96.5  | 33.21 | 93.46 |
| $Q_H$ value | 0.98  | 0.75  | 0.96  |

Structural alignment in different MD simulations

To investigate the structural similarities among the systems sampled in different pH conditions, we have aligned all the structures in pair using Multiseq plugin of VMD.
Figure S1: Superimposition of the high-resolution crystallographic structures of HCA IX (PDB id: 6FE2) with (a) 3IAI (b) 4ZAO and, (c) 5DVX. The blue areas indicate that the molecules are structurally conserved at those points. If there is no correspondence in structural proximities, the regions appear red.

Table S2: RMSD, percentage of sequence and structural similarity, and $Q_H$ value calculated for the superimposition of the structure pairs extracted from different MD simulation conditions.

|                     | HCA IX-c (CpHMD, pH 4.7 and 7.0) | HCA IX-c (classical MD and CpHMD, pH 7.0) | HCA II (classical MD and CpHMD, pH 7.0) | HCA IX-c and HCA II (classical MD) | HCA IX-c and HCA II (CpHMD, pH 7.0) |
|---------------------|----------------------------------|-----------------------------------------|---------------------------------------|-----------------------------------|-----------------------------------|
| RMSD (Å)            | 1.8430                           | 1.5259                                  | 2.2283                                | 2.2348                            | 2.4308                            |
| Percentage identity | 96.06                            | 84.98                                   | 83.14                                 | 31.6                              | 31.37                             |
| $Q_H$ value         | 0.7472                           | 0.8072                                  | 0.7171                                | 0.6454                            | 0.6018                            |

Structural stability in protein

The structural stability of the protein structure following a $1 \mu$s classical molecular dynamics simulation for HCA IX-c as well as HCA II has been defined in terms of the mass-weighted root mean square deviations (RMSD) and radius of gyration ($R_g$). The RMSD and $R_g$ plots of HCA IX-c highlighted in Figure S2(a,b) from MD and Figure S3(a,b) from CpHMD as well as those of HCA II highlighted in Figure S4(a,b) and Figure S5(a,b) from MD and CpHMD, respectively, clearly indicate the presence of a fairly stable structure. The $R_g$ value has been found to be varying from 17.3 Å to 17.8 Å for both HCA IX-c and HCA II in the case of classical MD and 17.3 Å to 17.8 Å for HCA IX-c and 17.5 Å to 18.0 Å for HCA II from the CpHMD trajectory. The root mean square fluctuations (RMSF) calculated for all the residues of HCA IX-c (Figure S2(c), S3(c)) and HCA II (Figure S4(c), S5(c)) showed no major secondary structure changes throughout the $1\mu$s classical MD simulation. Those residues with RMSF > 1.5 Å were mainly found to belong to the parts of loops, coils and
beta sheets on the protein surface. Lastly, the B-factors of $C_\alpha$ atoms of all amino acid residues have been calculated over $1 \mu s$ classical molecular dynamics simulation and have been compared to that of the crystallographic structure for HCA IX-c (Figure S2(d), S3(d)) as well as HCA II (Figure S4(d), S5(d)). An exact mapping of which was not observed owing to the finite temperature and solvent effects. The RMSF values thus obtained from the simulations correctly identify the flexible regions of the enzyme.

Figure S2: (a) Mass-weighted root mean squared deviation (RMSD) of backbone atoms with respect to energy minimised structure as a reference. (b) Radius of gyration ($R_g$) of the catalytic domain of HCA IX post $1 \mu s$ MD simulation run. (c) Root mean square fluctuations (RMSF) and (d) B-factor of $C_\alpha$ atom of all amino acid residues of the catalytic domain of HCA IX along $1 \mu s$ MD production run.

The corresponding variations of secondary structure elements of HCA IX-c present in their crystal structures and those in equilibrated structures at $pH = 4.5$ have been presented in Table S3. It is to be noted that even in the presence of such a highly acidic pH, the core of the protein remains almost intact with only minor changes to be observed.
Figure S3: (a) Mass-weighted root mean squared deviation (RMSD) of backbone atoms with respect to energy minimised structure as a reference. (b) Radius of gyration ($R_g$) of the catalytic domain of HCA IX along 1 µs CpHMD simulation run. (c) Root mean square fluctuations (RMSF) and (d) B-factor of $C\alpha$ atom of all amino acid residues of the catalytic domain of HCA IX along 1 µs CpHMD production run.

**Choice of secondary order parameters**

Now that the primary and secondary order parameters have been identified for both the systems, these can serve as the input for dimensionality-reduction methods. In this work, we have used time-structure independent components analysis (tICA) to carry out dimension reduction.

Towards this end, we have optimized the two slowest collective variables ($TIC_1$ and $TIC_2$) as a linear combination of the given input degrees of freedom. The relative weights of the OPs are the coefficients in $TIC_1$ and $TIC_2$.

As the $TIC_1$ preserves the maximum kinetic variance along the projection, it supposedly has more time-lagged correlations along the time series. Thus, it is generally assumed that $TIC_1$ takes more time to de-correlate than its other counterparts. For HCA II, the $TIC_1$ variable indeed takes more time to de-correlate when compared with $TIC_2$. However, for HCA IX-c, it is $TIC_2$ which is kinetically more sluggish. We have checked the individual
Figure S4: (a) Mass-weighted root mean squared deviation (RMSD) of backbone atoms with respect to energy minimised structure as a reference. (b) Radius of gyration ($R_g$) of all heavy atoms of HCA II along 1 $\mu$s MD simulation run. (c) Root mean square fluctuations (RMSF) and (d) B-factor of $C_\alpha$ atom of all amino acid residues in HCA II along 1 $\mu$s MD production run.

Figure S5: (a) Mass-weighted root mean squared deviation (RMSD) of backbone atoms with respect to energy minimised structure as a reference. (b) Radius of gyration ($R_g$) of all heavy atoms of HCA II along 1 $\mu$s CpHMD simulation run. (c) Root mean square fluctuations (RMSF) and (d) B-factor of $C_\alpha$ atom of all amino acid residues in HCA II along 1 $\mu$s CpHMD production run.
Table S3: Different segments and their associated major secondary structure element present in the equilibrated structures of HCA IX-c in water determined using STRIDE. The list of secondary structure elements have been prepared based on the crystal structure of HCA II (PDB id: 2ILI).

| Region | Major secondary structure | HCA IX-c |
|--------|---------------------------|----------|
|        |                           | PDB id: 6FE2^54 | pH=4.5 |
| 1.     | αA                        | 13-19     | 16-18   |
| 2.     | αB                        | 21-24     | 21-23   |
| 3.     | βa                        | 32-33     | 39-40   |
| 4.     | αC                        | 34-36     | -       |
| 5.     | βJ                        | 39-40     | -       |
| 6.     | βb                        | 46-50     | -       |
| 7.     | βB                        | 56-61     | 57-61   |
| 8.     | βC                        | 66-70     | 66-69   |
| 9.     | βc                        | 78-82     | -       |
| 10.    | βD                        | 87-97     | 88-97   |
| 11.    | βd                        | 108-109   | 108-109 |
| 12.    | βE                        | 116-124   | 116-124 |
| 13.    | αD                        | 131-135   | 131-135 |
| 14.    | βF                        | 141-150   | 141-150 |
| 15.    | αE                        | 155-162   | 155-166 |
| 16.    | βA                        | 172-178   | 172-176 |
| 17.    | αF                        | 181-184   | 181-184 |
| 18.    | βH                        | 191-196   | 191-197 |
| 19.    | βG                        | 207-212   | 205-212 |
| 20.    | βe                        | 216-218   | 216-218 |
| 21.    | αG                        | 220-227   | 220-227 |
| 22.    | βi                        | 257-258   | 257-258 |

OP de-correlation times for the two cases to address this issue. Interestingly, the TIC de-correlations are influenced by the mostly contributed OP de-correlations. For HCA II, TIC\(_1\) has maximum contribution coming from \(\chi_1\) and \(\chi_1\) also takes more time to de-correlate than the other OPs, making TIC\(_1\) to capture the maximum kinetic variance in input data. \(\chi_2\) takes more time to de-correlate in HCA IX-c, and it has a maximum contribution to TIC\(_2\), making the TIC\(_2\) variable kinetically slowest. The details of which are presented in the Figure S8.
Table S4: List of all candidates for primary and secondary OPs.

| Serial no. | CV      | Atoms involved                  |
|------------|---------|---------------------------------|
| 1          | $\chi_1$(Trp5) | N-C$_\alpha$-C$_\beta$-C$_\gamma$ |
| 2          | $\chi_2$(Trp5) | C$_\alpha$-C$_\beta$-C$_\gamma$-C$_\delta_2$ |
| 3          | $\Phi$(Trp5)  | C-N-C$_\alpha$-C               |
| 4          | $\Psi$(Trp5)  | N-C$_\alpha$-C-N               |
| 5          | $\chi_1$(Tyr7) | N-C$_\alpha$-C$_\beta$-C$_\gamma$ |
| 6          | $\chi_2$(Tyr7) | C$_\alpha$-C$_\beta$-C$_\gamma$-C$_\delta_1$ |
| 7          | $\Phi$(Tyr7)  | C-N-C$_\alpha$-C               |
| 8          | $\Psi$(Tyr7)  | N-C$_\alpha$-C-N               |
| 9          | $\chi_1$(Asn62) | N-C$_\alpha$-C$_\beta$-C$_\gamma$ |
| 10         | $\chi_2$(Asn62) | C$_\alpha$-C$_\beta$-C$_\gamma$-O$_\delta_1$ |
| 11         | $\Phi$(Asn62) | C-N-C$_\alpha$-C               |
| 12         | $\Psi$(Asn62) | N-C$_\alpha$-C-N               |
| 13         | $\chi_1$(His64) | N-C$_\alpha$-C$_\beta$-C$_\gamma$ |
| 14         | $\chi_2$(His64) | C$_\alpha$-C$_\beta$-C$_\gamma$-N$_\delta_1$ |
| 15         | $\Phi$(His64) | C-N-C$_\alpha$-C               |
| 16         | $\Psi$(His64) | N-C$_\alpha$-C-N               |
| 17         | $\chi_1$(Gln67) | N-C$_\alpha$-C$_\beta$-C$_\gamma$ |
| 18         | $\chi_2$(Gln67) | C$_\alpha$-C$_\beta$-C$_\gamma$-C$_\delta$ |
| 19         | $\Phi$(Gln67) | C-N-C$_\alpha$-C               |
| 20         | $\Psi$(Gln67) | N-C$_\alpha$-C-N               |
| 21         | $\chi_1$(Gln92) | N-C$_\alpha$-C$_\beta$-C$_\gamma$ |
| 22         | $\chi_2$(Gln92) | C$_\alpha$-C$_\beta$-C$_\gamma$-C$_\delta$ |
| 23         | $\Phi$(Gln92) | C-N-C$_\alpha$-C               |
| 24         | $\Psi$(Gln92) | N-C$_\alpha$-C-N               |
| 25         | $\chi_1$(Glu106) | N-C$_\alpha$-C$_\beta$-C$_\gamma$ |
| 26         | $\chi_2$(Glu106) | C$_\alpha$-C$_\beta$-C$_\gamma$-C$_\delta$ |
| 27         | $\Phi$(Glu106) | C-N-C$_\alpha$-C               |
| 28         | $\Psi$(Glu106) | N-C$_\alpha$-C-N               |
| 29         | $d_1$         | $Zn^{2+}$-N$_\delta_1$(His64)  |
| 30         | $d_2$         | Side chain O(Tyr7)-N$_\delta_1$(His64) |
| 31         | $d_3$         | C$_\delta_1$(Trp5)-N$_\delta_2$(Asn62) |
Figure S6: Population distribution of (a) $\chi_2$ of His-64, $d_1$, $d_2$, and $d_3$ as secondary order parameters for HCA IX-c from classical MD, (b) $\chi_2$ of His-64, $d_1$ and $d_2$ as secondary order parameters for HCA IX-c from CpHMD, (c) $d_1$ and $d_2$ as secondary order parameters for HCA II from classical MD, and (d) $d_1$ and $d_2$ as secondary order parameters for HCA II from CpHMD.
Figure S7: Correlation of input OPs to the TICs in the case of (a) HCA IX-c from classical MD, (b) HCA IX-c from CpHMD, (c) HCA II from classical MD and (d) HCA II from CpHMD.
Figure S8: The auto-correlation function of (a) TIC$_1$ and TIC$_2$, (b) individual OP in HCA IX-c using classical MD, (c),(d) for HCA IX-c using CpHMD, and (e), (f) for HCA II as a function of time.

Validation of MSM models

A Markov model is validated by estimating implied time scales (ITS) with $\tau$ being the lag time. Here, we look for the ITS convergence and choose the lag time accordingly, i.e., within a range where the ITS are approximately invariant. As shown in Figure S9, there are 4 implied timescales of the slowest processes for HCA IX-c obtained from classical MD as well as CpHMD trajectories and HCA II obtained from classical MD trajectory, whereas 5 implied timescales for HCA II derived from CpHMD trajectory. Nevertheless, we have seen that the time scales have reached a plateau region for Chapman-Kolmogorov (CK) test.
The Markov models that have been built at some specified lag times for the system were validated further using the Chapman–Kolmogorov test. The standard test computes the transition probability between metastable states for different lag times. The Chapman–Kolmogorov test was found to be sufficiently validated for all the systems (Figure S10).

References

(S1) Alterio, V.; Hilvo, M.; Di Fiore, A.; Supuran, C. T.; Pan, P.; Parkkila, S.; Scaloni, A.; Pastorek, J.; Pastorekova, S.; Pedone, C., et al. Crystal structure of the catalytic domain of the tumor-associated human carbonic anhydrase IX. Proc. Natl. Acad. Sci. U.S.A. 2009, 106, 16233–16238.
Figure S10: Model validation by Chapman–Kolmogorov test for (a) HCA IX-c (classical MD),
(b) HCA IX-c (CpHMD, pH 7.0), (c) HCA II (classical MD) and (d) HCA II (CpHMD, pH 7.0).

(S2) Pinard, M. A.; Aggarwal, M.; Mahon, B. P.; Tu, C.; McKenna, R. A sucrose-binding site provides a lead towards an isoform-specific inhibitor of the cancer-associated enzyme carbonic anhydrase IX. Acta Crystallogr., Sect. F: Struct. Biol. Commun. 2015, 71,
1352–1358.

(S3) Mahon, B. P.; Bhatt, A.; Socorro, L.; Driscoll, J. M.; Okoh, C.; Lomelino, C. L.; Mboge, M. Y.; Kurian, J. J.; Tu, C.; Agbandje-McKenna, M., et al. The Structure of Carbonic Anhydrase IX Is Adapted for Low-pH Catalysis. *Biochemistry* 2016, 55, 4642–4653.

(S4) Kazokaitė, J.; Niemans, R.; Dudutienė, V.; Becker, H. M.; Leitāns, J.; Zubrienė, A.; Baranauskienė, L.; Gondi, G.; Zeidler, R.; Matulienė, J., et al. Novel fluorinated carbonic anhydrase IX inhibitors reduce hypoxia-induced acidification and clonogenic survival of cancer cells. *Oncotarget* 2018, 9, 26800–26816.

(S5) Roberts, E.; Eargle, J.; Wright, D.; Luthey-Schulten, Z. MultiSeq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinf.* 2006, 7, 382.

(S6) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* 1996, 14, 33 – 38.

(S7) Eastwood, M. P.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. G. Evaluating protein structure-prediction schemes using energy landscape theory. *IBM J. Res. Dev.* 2001, 45, 475–497.

(S8) Heinig, M.; Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* 2004, 32, W500–W502.

(S9) Fisher, S. Z.; Maupin, C. M.; Budayova-Spano, M.; Govindasamy, L.; Tu, C.; Agbandje-McKenna, M.; Silverman, D. N.; Voth, G. A.; McKenna, R. Atomic crystal and molecular dynamics simulation structures of human carbonic anhydrase II: insights into the proton transfer mechanism. *Biochemistry* 2007, 46, 2930–2937.