CICLOP: A Robust, Faster, and Accurate Computational Framework for Protein Inner Cavity Detection.

Parth Garg [1], Sukriti Sacher [1], Arjun Ray [1] *

[1]Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla, India

Abstract

Internal cavities of proteins are of critical functional importance. Yet, there is a paucity of computational tools that can accurately, and reliably characterize the inner cavities of the proteins, a prerequisite for elucidating their functions. Here we introduce CICLOP a novel computational tool that accurately, reproducibly, and speedily identifies residues lining the inner cavity of a protein, its morphometric dimensions, hydrophobicity as well as evolutionary significance.

Proteins are the major drivers of diverse cellular processes. The functionality of these bio-molecular machines largely depends on their three-dimensional structure, accurate estimation of which is an ever-challenging task. For several proteins, presence of topological features such as clefts, grooves, protrusions, and internal cavities further increase their structural complexities. These features serve as ligand binding sites [1;2], active or allosteric sites in enzymes [3], channels for transportation of small molecules [4] and sometimes just niche environments excluded from the bulk solvent [5;6]. Of note, protein cavities, one of the functionally indispensable features, mediate the conformational changes occurring between domains or subunit interfaces during structural transitions [7;8]. Additionally, the molecular composition and the physicochemical properties of these cavities is known to impart specificity and selectivity towards their cognate biomolecule [9;10]. Hence, identification and characterization of tunnels/channels is of immense importance in order to deduce their function.

In pursuit of this, various methods [11;12;13;14;15] have been proposed. Unfortunately, many are limited in their functionality, accuracy, automation, and comprehensiveness, while some even require user intervention and advanced knowledge about the protein. We have developed CICLOP (Characterization of Inner Cavity Lining Of Proteins), an end-to-end automated solution for the identification and characterization of protein cavities at an atomic resolution. CICLOP builds on a novel algorithm that imparts unprecedented speed, accuracy, and reproducibility, outperforming its predecessors. We have implemented the method as a webserver, allowing users to perform an in-depth analysis by merely uploading the protein structure file (PDB format). Supplementary T1 summarizes the features and strengths of our tool in comparison to the other leading methods.

In automatic mode, the algorithm rotates the input structure such that its central pore axis lies along the Z-axis while in case of manual operation, the same is assumed. Using the input of the PDB three dimensional file format, CICLOP maps the protein structure to a 3D grid consisting of cubes, where each cube is treated as a node in a directed cyclic graph. Our algorithm then performs a breadth first search to find all the continuously empty regions taking any random cube as the starting point. Subsequently, numerous thin "slices" along the central pore axis are cut and the lining of the cavity is elucidated by calculating the statistical mean and standard deviation of distances of the atoms detected in the initial search from the geometrical center of the protein (See M&M for details). The final output is a B-factor loaded PDB file marking the residues detected on the inner surface of the cavity. Furthermore, using vertices given by voronoi diagrams to be the centre of the circle, the diameter for a slice is estimated as the largest disk that can fit in the region enclosed by atoms detected on the inside. Finally, the estimation of the total pore volume is performed using the sum of the areas enclosed by all the inner lining atoms (Supplementary Fig 1 and 2).

CICLOP includes several analysis modules, aiding in the functional characterization of cavities. Conserved regions in proteins often point towards a functional domain that either confer structural stability or serve

*To whom correspondence should be addressed. Tel: +91 11 26907438; Fax: +91 11 2690 7405; Email: arjun@iiitd.ac.in

as active sites[16]. CICLOP allows for computation of conservation scores of cavity-lining residues, which are normalized for comparison between proteins. Furthermore, a detailed profile of secondary structure and charge distribution is provided in the output summary file. Figure 1 highlights the various features and strengths of our tool. Our tool successfully identified residues lining the inner cavity of human mitochondrial chaperonin (Figure 1a). The conservation module was also able to recapitulate that 57.836/16.592 % of the residues lining the cavity of this protein are highly/moderately conserved, while those lining the two "caps" are less conserved (Figure 1b). Such features help in understanding evolutionary pressure exerted at different sites of a complex structure. The analysis highlights that the cavity-lining lesser conserved residues have a larger propensity to exist as turns (14.11%), in comparison to the highly conserved residues (3.16%) (Figure 1c). Additionally, conservation classification of the inner-residues as a function of the cavity length, along the Z-axis is also provided (Figure 1d). The intricate volume and diameter profile clearly capitulates the interface of the upper and lower cavity at $\approx$ 125 Å (Figure 1e,f) (Supplementary Fig 2). Our tool can also be used to detect cavities at the interface of multimers such as those formed by a homo-trimeric arrangement in spike protein of SARS-CoV2 virus (Figure 1g).

CICLOP was tested against a diverse set of cavity morphology as well as proteins with varied sub-cellular localizations (Supplementary Fig 3). CICLOP's atomistic resolution also aided in identifying a gradient of residue accessibility of the inner residues versus the rest of the molecule (Supplementary Fig 4). We also tested the performance of our method against several leading cavity detection methods using a set of proteins that varied in their geometric shapes. The inner residues lining the cavity, as calculated by CICLOP along with two other leading methods – PoreWalker and MoleOnline have been highlighted (Supplementary Fig 5, 6). The robustness of our tool was tested on a massive protein complex of human parechovirus (HPeV) epitope containing 302,100 atoms arising from 38,580 residues in its four unique chains. In comparison to any of the previous methods, which either were unable to process the file or gave inaccurate results, CICLOP was able to automatically analyse the structure, without any glitches (Figure1h,i).

As a measure of speed, we plotted the computation time as a function of the size of protein taken by various tools (Supplementary Fig 7) and observed that CICLOP consistently outperformed by many orders of magnitude (Supplementary T5). To understand the accuracy and precision of various methods, we performed all-atomistic molecular dynamic simulations of representative proteins (PDBID:1TF7,6V0B,1AON), in order to identify the water accessible cavity residues. Residues identified to be on the inner surface (See M&M for details) in this simulation were then compared to the list of residues detected by CICLOP and other tools – PoreWalker, MoleOnline and CaverWeb. Our tool was able to detect the inner residues lining the cavity with an unparalleled accuracy (85.22–91.52%) and precision (90.01–99.15%), compared to the sub-optimal performance of other tools (Supplementary Fig 8-10, table 2-4).

In order to demonstrate our tool's applicability, we employed CICLOP to characterize the cavity of the F1 domain of bovine mitochondrial ATP synthase, as a case study. ATP synthases are found in the inner membrane of mitochondria and operate by a rotary catalytic mechanism. This highly conserved biomotor functions by coupling proton translocation through the F0 domain to the rotation of a central rotor ($\gamma\delta\varepsilon$) in the F1 complex ($\alpha\beta\gamma\delta\varepsilon$)(Figure 2a), generating ATP in the process[17]. Using Cryo-EM, Zhou et al. obtained three rotational states of bovine ATP synthase that were related to each other by a rotation of 120°. Each of these states were further divided into seven sub-states providing a snapshot of ATP synthase during its full catalytic cycle[18]. We used these seven representative snapshots as an input for CICLOP to deduce if our tool could detect the minute changes occurring in the cavity as the central rotor rotated about its axis. The detection sensitivity of our tool is evident in the comparison of the diameter profiles of the seven sub-states (Supplementary Fig 11). We further validated the diameter profile of State 1A by manually measuring the distance between opposite ends of the cavity (Supplementary Fig 12,13).

As the γ subunit complex rotates inside the cavity, it orients itself towards the interface of an α β subunit (Supplementary Fig 14). The cavity remains immobile, held in place by a peripheral stalk that connects it to the membrane embedded region of F0[19]. The motion of the rotor however, leads to a slight bobbing of the cavity about its axis[18]. This also results in conformational changes in the internal face of the cavity it orients towards[20] characteristic of nucleotide binding states. Firstly, to capture the minor physico-chemical perturbations arising during each state-change, we characterized the total hydrophobicity of the pore (Figure 2b). Additionally, the subtle structural variations amongst substates were evidently characterized in both diameter profile and pore volume calculated by CICLOP (Figure 2c-f). We further quantified the conformational changes occurring in the cavity during transition from one sub-state to another (Supplementary T6) with the help of a list of residues that line the cavity identified by our tool. The conformational changes occurring during each state transition is a result of movement of residues in α β chains that lie diagonally across each other (Supplementary Fig 15), while major movement is observed in chains towards which γ is oriented (Figure 2g and Supplementary Fig 14, 16). Moreover, during a rotary cycle, dynamic as well as constant residues are concentric to each other(Figure 2h and Supplementary Fig 17). We also employed the conservation and secondary structure module included in CICLOP and found that the internal organization of the cavity is highly conserved across all species (Supplementary Fig 18, 19) and the secondary structure distribution of the inner cavity residues

largely remains constant (Supplementary Fig 20).

In summary, we have demonstrated our tool's ability to quantitatively as well as qualitatively characterize the internal cavity of proteins. The analyses provided by CICLOP are more sensitive, precise and accurate than those provided by the current cavity detection methods. We have also demonstrated the use of various modules offered by our tool and how they can facilitate inference of biological functions. Internal cavities, although a predominant topological feature of proteins, to this day, remain elusive in terms of their functions. We expect that CICLOP will be useful in changing this paradigm by allowing an increased access to these otherwise isolated pockets in a protein structure. The atomistic detail provided by CICLOP will have major application in the field of structural biology in evaluating protein structures, identifying solvent accessible surfaces, protein oligomerization, understanding chaperone assisted protein folding as well as functional characterization of channels. The method is available at https://ciclop.raylab.iiitd.edu.in.

Acknowledgements

The authors would also like to thank HPC facility of IIIT Delhi for the computational facility. S.S was supported by CSIR-DBT funding agency. The study was supported by the Initiation Research Grant by IIIT Delhi for A.R.

Author contributions

P.G. and A.R designed the study and developed the algorithm. P.G. carried out all the programming, implementation of the algorithm and parameterizations. P.G. and S.S. analyzed the application of the method and P.G., S.S. and A.R. wrote the manuscript.

The authors declare no competing interests.

References

[1] Jan M Kriegl, Karin Nienhaus, Pengchi Deng, Jochen Fuchs, and G Ulrich Nienhaus. Ligand dynamics in a protein internal cavity. *Proceedings of the National Academy of Sciences*, 100(12):7069–7074, 2003.

[2] Lukáš Pravda, Karel Berka, Radka Svobodová Vařeková, David Sehnal, Pavel Banáš, Roman A Laskowski, Jaroslav Koča, and Michal Otyepka. Anatomy of enzyme channels. *BMC bioinformatics*, 15(1):379, 2014.

[3] Ryo Kitahara, Yuichi Yoshimura, Mengjun Xue, Tomoshi Kameda, and Frans AA Mulder. Detecting o 2 binding sites in protein cavities. *Scientific reports*, 6(1):1–12, 2016.

[4] Ajay Singh Tanwar, Venuka Durani Goyal, Deepanshu Choudhary, Santosh Panjikar, and Ruchi Anand. Importance of hydrophobic cavities in allosteric regulation of formylglycinamide synthetase: insight from xenon trapping and statistical coupling analysis. *PLoS One*, 8(11):e77781, 2013.

[5] Brian W Matthews and Lijun Liu. A review about nothing: are apolar cavities in proteins really empty? *Protein Science*, 18(3):494–502, 2009.

[6] Maurizio Brunori, Beatrice Vallone, Francesca Cutruzzolà, Carlo Travaglini-Allocatelli, Joel Berendzen, Kelvin Chu, Robert M Sweet, and Ilme Schlichting. The role of cavities in protein dynamics: crystal structure of a photolytic intermediate of a mutant myoglobin. *Proceedings of the National Academy of Sciences*, 97(5):2058–2063, 2000.

[7] Mengjun Xue, Takuro Wakamoto, Camilla Kejlberg, Yuichi Yoshimura, Tania Aaquist Nielsen, Michael Wulff Risør, Kristian Wejse Sanggaard, Ryo Kitahara, and Frans AA Mulder. How internal cavities destabilize a protein. *Proceedings of the National Academy of Sciences*, 116(42):21031–21036, 2019.

[8] Shrihari Sonavane and Pinak Chakrabarti. Cavities and atomic packing in protein structures and interfaces. *PLoS Comput Biol*, 4(9):e1000188, 2008.

[9] Peter Agre, Dennis Brown, and Søren Nielsen. Aquaporin water channels: unanswered questions and unresolved controversies. *Current opinion in cell biology*, 7(4):472–483, 1995.

[10] Declan A Doyle, Joao Morais Cabral, Richard A Pfuetzner, Anling Kuo, Jacqueline M Gulbis, Steven L Cohen, Brian T Chait, and Roderick MacKinnon. The structure of the potassium channel: molecular basis of k+ conduction and selectivity. *science*, 280(5360):69–77, 1998.

[11] Oliver S Smart, Joseph G Neduvelil, Xiaonan Wang, BA Wallace, and Mark SP Sansom. Hole: a program for the analysis of the pore dimensions of ion channel structural models. *Journal of molecular graphics*, 14 (6):354–360, 1996.

[12] Jan Stourac, Ondrej Vavra, Piia Kokkonen, Jiri Filipovic, Gaspar Pinto, Jan Brezovsky, Jiri Damborsky, and David Bednar. Caver web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport. *Nucleic acids research*, 47(W1):W414–W422, 2019.

[13] Lukáš Pravda, David Sehnal, Dominik Toušek, Veronika Navrátilová, Václav Bazgier, Karel Berka, Radka Svobodová Vařeková, Jaroslav Koča, and Michal Otyepka. Moleonline: a web-based tool for analyzing channels, tunnels and pores (2018 update). *Nucleic acids research*, 46(W1):W368–W373, 2018.

[14] Ryan G Coleman and Kim A Sharp. Finding and characterizing tunnels in macromolecules with application to ion channels and pores. *Biophysical journal*, 96(2):632–645, 2009.

[15] Eitan Yaffe, Dan Fishelovitch, Haim J Wolfson, Dan Halperin, and Ruth Nussinov. Molaxis: a server for identification of channels in macromolecules. *Nucleic Acids Research*, 36(suppl_2):W210–W215, 2008.

[16] Alejandro Panjkovich and Xavier Daura. Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. *BMC structural biology*, 10(1):1–14, 2010.

[17] Paul D Boyer. The atp synthase—a splendid molecular machine. *Annual review of biochemistry*, 66(1): 717–749, 1997.

[18] Anna Zhou, Alexis Rohou, Daniel G Schep, John V Bason, Martin G Montgomery, John E Walker, Nikolaus Grigorieff, and John L Rubinstein. Structure and conformational states of the bovine mitochondrial atp synthase by cryo-em. *Elife*, 4:e10180, 2015.

[19] John E Walker and Veronica Kane Dickson. The peripheral stalk of the mitochondrial atp synthase. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1757(5-6):286–296, 2006.

[20] Jan Pieter Abrahams, Andrew GW Leslie, René Lutter, and John E Walker. Structure at 2.8 a resolution of f1-atpase from bovine heart mitochondria. *Nature*, 370(6491):621–628, 1994.

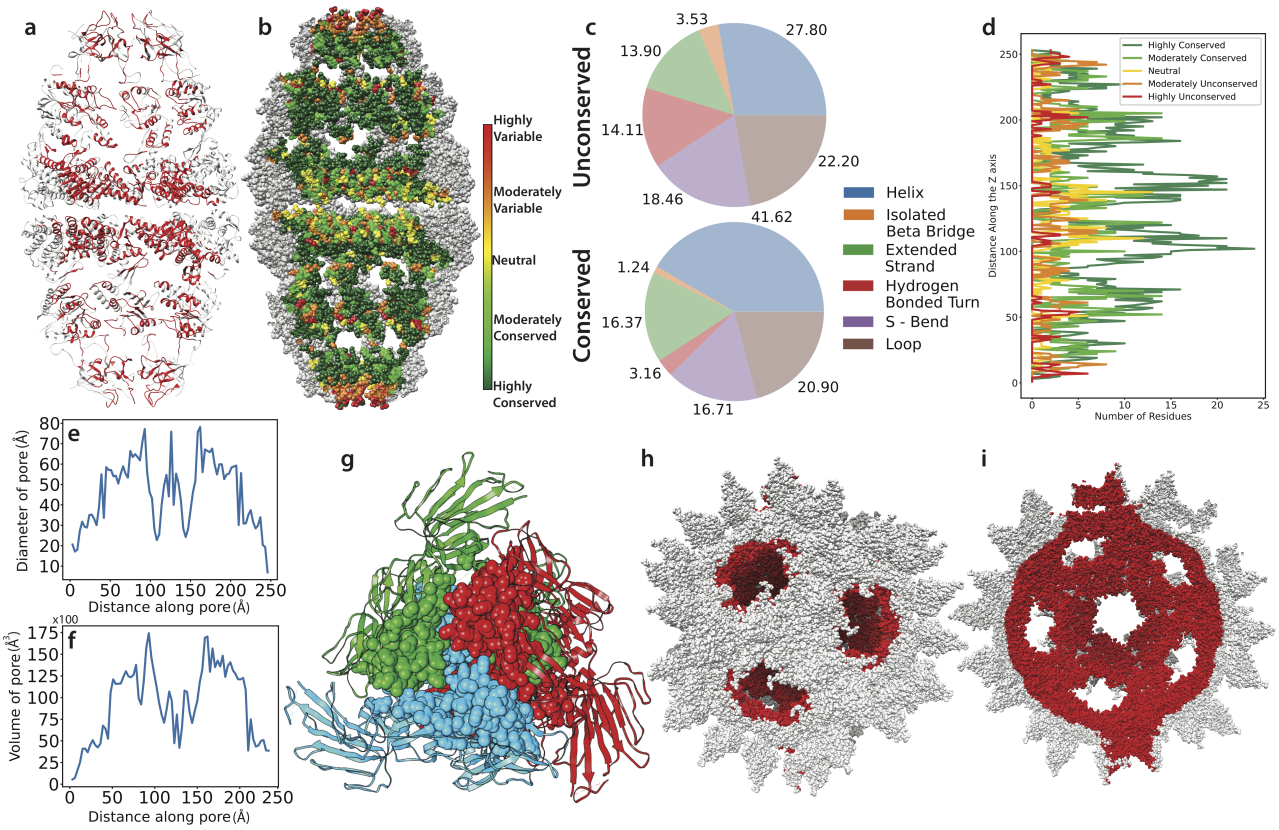

Figure 1: **Application of CICLOP:** a. Identification of the residues lining the inner surface of WT human mitochondrial chaperonin (ADP:BeF3)14 complex (PDBID: 6HT7). b. Conservation of the identified residues; marked in the range of highly variable (red) to highly conserved (dark green). c. Secondary structure assignment of the conserved and unconserved residues as detected by CICLOP along with d. Conservation profile as a function of z-axis e. Diameter profile of the pore of the protein as a function of z-axis distance. f. Volume profile of 6HT7 as a function of z-axis distance (Total pore volume = 738,319.197 $Å^3$). g. Top view of the oligomerization interface residues (forming a cavity), of the three chains (highlighted in red, blue and green) of the Alpha-coronavirus spike glycoprotein (PDBID: 6IXA), as detected by CICLOP and represented in spheres. h. Outer surface of Human Parechovirus (HPeV) protein complex (PDBID: 4UDF) with inner surface marked red and i. longitudinal section of the inner surface (red) as detected by CICLOP.

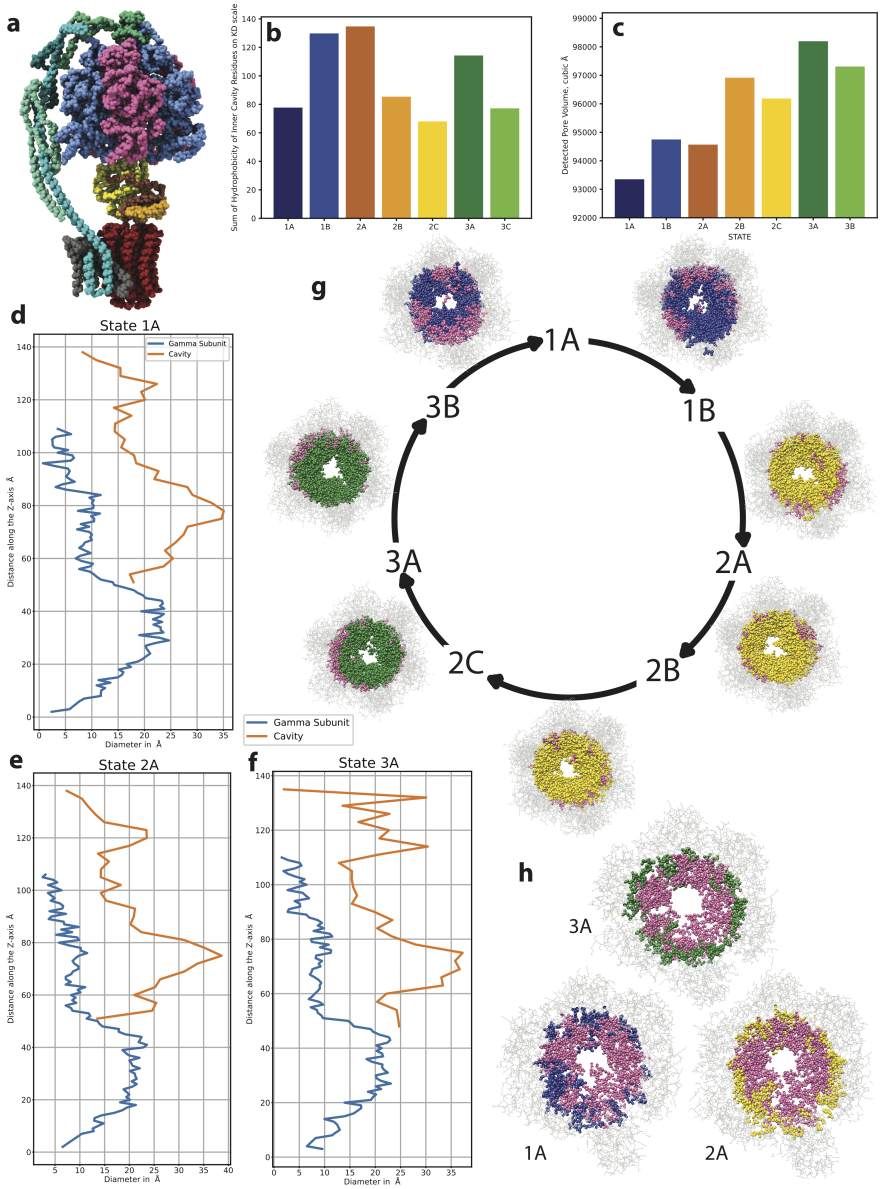

Figure 2: **Analysis of the internal cavity of ATP synthase F1 domain during its catalytic rotary cycle using CICLOP.** a. Structure of bovine mitochondrial ATP synthase (PDBID: 5ARA). Chains A-C(blue) that constitute the α subunits as well as chains D-F(pink) that constitute the β subunits are alternatively arranged to form the F1 cavity. Chain G (yellow), H (orange) and I (brown) that form the subunit γ, δ, ϵ respectively, together constitute the rotor complex of the F1 domain. Chains J-Q(red) that form the c1 subunit complex of F0 domain are inserted in the mitochondrial membrane. a1 subunit formed by chain W (gray) is attached to c1. The peripheral stalk formed by OSCP complex (chain S), subunit b1 (chain T), subunit d (chain U), subunit f6 (chain V) are colored as sea green, turquoise, bright blue and teal respectively. b. The overall hydrophobicity of the ATP synthase cavity during its rotation. Each state is represented by a differently colored bar. c. The pore volume of ATP synthase cavity during a complete rotation. d. The diameter profile generated using CICLOP of the ATP synthase cavity for State 1A. e. State 2A f. State 3A. g. Top view of the cavity during each state transition of the catalytic rotation. Residues facing the cavity as detected by CICLOP are indicated in blue, yellow and green colors respectively for state 1, 2 and 3 respectively. Dynamic residues for each sub state transition are marked in pink. h. Top view of the cavity of each representative state during the ATP synthase rotary cycle. Residues facing the cavity as detected by CICLOP are colored as blue, yellow and green for each state respectively while residues that are common in all the substates are colored as pink.

CICLOP: A Robust, Faster, and Accurate Computational Framework for Protein Inner Cavity Detection. - Methods Section

Parth Garg [1], Sukriti Sacher [1], Arjun Ray [1] *

[1]Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla, India

Materials and Methods

CICLOP maps and identifies the different residues that are accessible to ligands from the inner surface of any given protein. CICLOP is currently available as a webserver at the url https://ciclop.raylab.iiitd.edu.in. Each job on the server mandates the user to provide their email address along with the PDB file(s) of the proteins which the user wishes to characterize. This section describes the workings of the various modules included in our tool.

Alignment of the structure

The user is offered two choices for alignment of their protein structure. In the automatic mode, the protein structure is aligned based on the best-fit line that describes the positions of all the alpha carbon atoms. The protein is then rotated such that this best fit line lies along the positive Z-axis. In manual alignment, CICLOP skips the process of alignment of the protein and assumes that the provided structure has the best-fit line (or the central pore axis) aligned with either the negative or positive Z axis.

Identification and mapping of the Inner Surface Residues

The identification and mapping of inner surface residues by CICLOP is a six stage process.

Stage I: All the relevant information (such as coordinates of all atoms, the residues and protein chains) is extracted from the given PDB file. The structure is subjected to linear transformation such that the absolute minimum value from each of the x,y,z coordinate list is added to the respective x, y, z coordinates of all atoms comprising the protein. This places the protein in the first quadrant of the cartesian plane with coordinates of all atoms having either a positive value or a zero value. A box enclosing the protein is then defined.

Algorithm 1 Creation of Bounding Box

```
minX ← min(allXcoordinates)
minY ← min(allYcoordinates)
minZ ← min(allZcoordinates)
for all atoms do
    atom.xcoord ← atom.xcoord + minX
    atom.ycoord ← atom.ycoord + minY
    atom.zcoord ← atom.zcoord + minZ
BoxCoords = {0,0,0, max(all X coordinates), max(all X coordinates), max(all X coordinates)}
```

Stage II: The protein after the linear transformation is mapped to a 3-Dimensional grid of cubes similar to voxels (A voxel is a unit of graphic information that defines a point in three-dimensional space. Each constituent voxel cube has a volume of 1 $Å^3$. The dimensions of the enclosing box are then defined forming a voxel grid. All the protein atoms from the PDB structure are mapped onto a cube (from the complete voxel grid) on the basis of their coordinates. It may be noted that each cube in the 3D grid may either contain no atom, one atom or more than one atom. The cubes containing no atoms are identified and marked empty. Since, each voxel cube from the complete grid contains references to cubes bounding it from all of the six faces, therefore, each

*To whom correspondence should be addressed. Tel: +91 11 26907438; Fax: +91 11 2690 7405; Email: arjun@iiitd.ac.in

voxel also contains information about voxels lying on its boundaries allowing this data to be represented as a directed graph. Each node in this graph contains all the atoms lying in its vicinity and each edge represents all the atoms that bound the given node from any of the given six directions.

Stage III: A breadth first search algorithm is applied to look for empty "nodes" starting from the node containing coordinates of the geometrical center of the transformed protein. If a node is empty, it may be part of a cavity inside the protein, therefore, edges emanating from this node are traversed. The search continues in all six cardinal directions until no more empty nodes are encountered. This region physically represents the boundary of the protein (both, the inner and the outer). The atoms surrounding the identified empty regions are initially marked to be part of the inner surface of the protein. This is done to separate them from the bulk of the protein.

Stage IV: The next step involves separating the inner surface from the outer surface. The algorithm for this step involves making 1 Å slices of the protein along the Z axis. For each and every such slice made:

- The atoms marked to be on the inner surface are identified.
- The mean radius at which these residues lie from the geometrical centre of these marked atoms is identified.

$$X_{cen} = \frac{\Sigma atom.xcoord}{Total\ number\ of\ inner\ atoms}$$

$$Y_{cen} = \frac{\Sigma atom.ycoord}{Total\ number\ of\ inner\ atoms}$$

$$R_{mean} = \frac{\Sigma\sqrt{(X_i - X_{cen})^2 + (Y_i - Y_{cen})^2}}{Total\ number\ of\ inner\ atoms}$$

  Where X_{cen} and Y_{cen} are the X and Y coordinates of the geometric center of the inner atoms identified for the given Z slice. R_{mean} is the mean distance at which these inner atoms lie around the given Z coordinate

- The standard deviation of all the atoms (lying at a distance less than the R_{mean} from the geometrical centre) is calculated from the R_{mean}.

$$R_i = \sqrt{(X_i - X_{cen})^2 + (Y_i - Y_{cen})^2}$$

$$SD = \sqrt{\frac{\Sigma(R_i - R_{mean})^2}{Number\ of\ Inner\ atoms\ having\ R_i\ less\ than\ R_{mean}}}$$

  Where R_i is the distance from (X_{cen}, Y_{cen}) along the XY plane, R_{mean} is the mean radius and SD is the Standard Deviation of the R_i from R_{mean} $\forall$ atoms having $R_i \leq R_{mean}$

- Any atom lying outside the circle of $Radius = R_{mean}$ - (0.7) x SD is unmarked and removed from the list of atoms lying on the inside lining

Stages ***II***, ***III*** and ***IV*** are repeated until all the empty cubes have been traversed at least once.

Stage V: Once all the atoms lining the inner surface have been identified, a PDB file is rewritten with the original coordinates and a new B factor value such that the temperature factor 9999 is assigned to atoms lying on the inner surface while a value of 0 is allotted to all others. This ensures that the user can effectively visualize the atoms lying on the inner surface identified by CICLOP.

Stage VI: Finally residues to which these atoms belong to, is written to a separate file called 'residue.dat' and is provided as one of the outputs. The list of residues that line the surface contains all those residues for which a single atom was detected to be on the inner surface. Considering these calculations are perfomed on static structures, even if a single atom of a residue is detected on the inner lining, under real life dynamic situations, it is likely that the complete residue becomes available for interactions in the cavity.

Calculation of the Pore Diameter Profile

This module is helpful in identifying other biomolecules/ions that can pass through the pore at a given point in space aiding in structural analysis of a pore/channel protein. CICLOP automatically generates a diameter profile for all the jobs submitted. The calculation of the pore diameter is based on the largest circle that can fit inside the pore at the given length along the pore axis. The diameter measurements are taken at every 3 Å intervals along the pore axis. This cut off was chosen because most of the atoms of a protein have a mean Van-der-Waal diameter of 3.6 Å[1], and therefore, the diameter is unlikely to change at a smaller step size.

The pore diameter is calculated by calculating the voronoi diagram of all the atoms that are identified on the inner surface in the 3 Å block in order to obtain its voronoi vertices.

If X is a metric space with a distance function d. Let K be a set of indices and $(P_k)_{k \in K}$ be an ordered collection of non-empty sites in the space X. The Voronoi Region (R_k), associated with the site P_k is the set of all points in X whose distance to P_k is not greater than their distance to other sites in X.

$$R_k = \{x \in X \mid d(x, P_k) \leq d(x, P_j) \textit{ for all } j \neq k\}$$

Voronoi-Vertices are the points where 3 or more of the vornoi regions (R_k) intersect.
The distance used in our study is the familiar *Euclidean Distance* that is calculated as follows:

$$d[(x1, y1), (x2, y2)] = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

Similarly, a convex hull is calculated for all the atoms lying on the inner surface in the same block.

Formally, the convex hull for a collection of points is the minimum n-sided convex polygon which completely encloses the given collection of points.

Let a simple polygon have n vertices x_i for $i = 1, 2, 3..., n$ and define the edge vectors as

$$v_i = x_{i+1} - x_i,$$

where x_{n+1} is understood to be equivalent to x_1. Then the polygon is convex if and only if all turns from one edge vector to the next have the same sense. Therefore a simple polygon is convex if,

$$v_i^{\perp}.v_{i+1}$$

has the same sign for all i where $a^{\perp}.b$ denotes the perpendicular dot product of two vectors.

The voronoi vertices that lie inside this calculated convex hull are marked as possible centers. For all such centers, the distance between the innermost surface atom which is closest to the center is identified and this distance is marked as a possible radii. The maximum value of all such radii is doubled and 3 Å is subtracted from that value to account for Van-der-Wall radii from the opposite ends. The value so obtained is the diameter for the given block. This process is repeated for all the 3 Å blocks and plotted as a diameter along the Z axis plot, provided as an output to the user.

Calculation of the Pore Volume Profile

The measure of volume is an estimate of the largeness of the pore/cavity and is another helpful structural feature. CICLOP automatically generates a volume profile of all the jobs submitted. The volume is calculated for every 1 Å interval along the pore axis, however, similar to the diameters, it is recorded for every 3 Å block along the pore axis. Total volume of a block is the sum of the volumes of the three individual 1 Å slices cut perpendicular to the pore axis that make up the block.

For every 1 Å slice, its pore volume value is obtained by first calculating the geometric center of all the atoms lining the inner surface of that slice. Taking the center as the reference point, this list of atoms is sorted in a clockwise direction. On joining every unique pair of consecutive points with the center, a triangle is formed and the area of all such triangles is calculated and summated. The division of the entire circular slice into smaller triangles that all have the center of the circle as a common vertex ensures that no two triangles have overlapping areas. The volume of the slice is measured by multiplying the height of slice (1 Å in this case) with the total area of the slice.

The area for each triangular region is calculated using the widely used *Heron's Formula*. For a triangle formed using 3 points P_1 P_2 P_3 such that P_i is the point (x_i, y_i), the area enclosed by the triangle is given by

$$\Delta = \sqrt{s(s-a)(s-b)(s-c)}$$

Where the meaning of the symbols is as follows

$$a = d[(x_1, y_1), (x_2, y_2)]$$
$$b = d[(x_2, y_2), (x_3, y_3)]$$
$$c = d[(x_3, y_3), (x_1, y_1)]$$
$$s = \frac{a+b+c}{2}$$

Where $d[(x_i, y_i), (x_j, y_j)])$ is the euclidean distance between (x_i, y_i) and (x_j, y_j)

Creation of the Result Summary File

CICLOP generates a summary file for each job submitted in addition to the pore volume, radius and conservation profiles as well as inner surface residues detected as described previously. This summary contains the number of residues and atoms marked to be on the inner surface of the protein as detected by CICLOP, the total pore volume as well as the sum of the hydrophobicities (on the Kyte-Doolittle scale[2]) of all the inner surface residues. Secondary structure information extracted using DSSP[3] and mapped to the crystal structure is also provided in the form of the number of inner surface residues that are part of the various secondary structures in the summary file. The total number of charged amino acids as well as the number of positively charged and negatively charged residues lying on the inside of the cavity is also included in this file. The time taken by CICLOP (in seconds) to calculate the inner surface and perform the relevant analysis is also included in the summary. The time taken by the tool for the calculation of the conservation scores (in case the user has opted for the same) is not included in this calculation.

Evaluating Residue Conservation and Calculation of the Conservation Scores

This module, unique to CICLOP, is helpful in identifying regions lining the inner cavity that may be structurally or functionally relevant to the cavity. For the estimation of conservation percentage of amino acid residues lining the cavity, first all the unique chains from the input PDB file are extracted. A basic local alignment search is performed on each unique chain in the PDB file against a local copy of nr-database of proteins. Other protein sequences similar to the reference sequence are found using the PSI-BLAST package from the NCBI BLAST suite[4]. A multiple sequence alignment using Muscle[5] is performed amongst the reference sequence as well as the other similar sequences obtained from PSI-BLAST. The rate4site method[6] is applied to this MSA file to calculate the evolutionary scores.

These conservation scores, so obtained form an almost continuous distribution. To aid in interpretation and improve visualization, these are therefore grouped into 9 distinct categories. The method to do so is similar to the method used by CONSURF[7]. An average of all the conservation score values is calculated and the scores below as well as scores above the average are are both divided into 4.5 equal intervals each (resulting in 9 equally sized categories of conservation).

It may be noted that the width of each bin generated using the given procedure may vary from one polypeptide to another. It does not indicate the absolute magnitude of evolutionary distances, but rather is a measure of the relative degree of conservation of each amino acid position[7]. These conservation scores are mapped onto the inner surface atoms of the protein. Another PDB file is written with the original coordinates and the conservation scores of all atoms lining the inner surface written as b factor values. For all the atoms not on the inner surface, the b-factor takes the value of 10.

After successfully mapping the conservation scores calculated onto the inner lying residues, CICLOP moves through the structure along the Z axis. The residues lying on the inner surface in each Å step are then divided into one of the 5 groups based on the conservation score assigned to them (Highly Conserved(1-2), Moderately Conserved(3-4), Neutral(5), Moderately Unconserved(6-7), Highly Unconserved(8-9)). This is represented as a color coded plot with the number of residues on the X axis and the distance along the cavity on the Y axis.

Calculation of conservation scores is an optional module. If chosen, the user is mandated to provide the following parameters :

(a) ***BLAST-Evalue:*** This is the number of expected hits of similar quality (score) that can be found just by chance. This value is set at 10 by default and can be changed by the user.

(b) ***Method for estimating evolutionary rate (two rate inference):***The two methods that can be chosen are described as follows

- ***Empirical Bayesian:*** This method assumes a prior gamma distribution over the evolutionary rates. Bayesian inference is based upon the posterior probability distribution, which is directly proportional to the product of the prior distribution and the likelihood. This method is considered to be superior to the maximum likelihood method of rate determination[8]
- ***Maximum likelihood:*** This method makes no presumptions regarding the prior distributions of evolutionary rates[8]

(c) ***Evolutionary model:*** The seven substitution models offered by CICLOP used to infer the evolutionary conservation scores are as follows:

- ***JC amino acids:*** This model assumes that amino acids occur in equal frequency and are substituted in equal probability[9]. This method is chosen by default.

- ***Dayhoff Model:*** The empirical model of amino acid substitution developed by Dayhoff, Schwartz and Orcutt(1978) assumes that all sites in a protein evolve independently of one another. The process of amino-acid replacement, at each site is defined by a matrix of replacement rates. All such sites in a protein according to this model evolve according to the same rate matrix[10].
- ***JTT Model:*** A new and improved version of the Dayhoff model developed by Jones, Taylor and Thornton using a much larger dataset to construct a replacement matrix as proposed in the Dayhoff model[10].
- ***REV Model:*** The is the general reversible Markov process model of amino acid substitution. It places a constraint on the structure of the rate matrix. For this model, the probability of the first amino acid changing to the second is kept the same as the probability of the second amino acid changing to first[10].
- ***WAG Model:*** A comparatively new model, offering a better statistical fit to the data. This method may provide more accurate estimates of phylogenetic trees than other existing models[11].
- ***LG Model:*** This model is an improvement to the WAG model, by incorporating the variability of evolutionary rates across sites in the matrix estimation and using a larger and more diverse database than BRKALN which was used to estimate the WAG model[12].
- ***cpREV Model:*** A model similar to the REV model but optimised for chloroplast proteins[13].

Based on the parameters provided above, and after the completion of the submitted job, the results of the PDB file(s) are sent to the email address supplied by the user as a downloadable link. Additionally, the web server includes a section dedicated to interpretation of results which illustrates how the output files can be visualized. We also provide the scripts that integrate with commonly used molecular visualizers, along with the instructions, to further aid in visualization of inner surface residues. These scripts are freely available in the "Downloads" section of the web server.

Collection of Data Used in this Study

All the protein structures used in this study were obtained from the RCSB database and their PDB IDs have been indicated in their respective figures. The structures used in the ATP synthase case study were as follows: 5ARA (State 1A), 5ARE (State 1B), 5ARH (State 2A), 5ARI (State 2B), 5FIJ (State 2C), 5FIK (State 3A), 5FIL (State 3B). All the images used in this study were produced using UCSF Chimera[14] (except the full structure of ATP synthase, which was produced using Chimera X[15].)

Identification of Inner Residues Using Various Tools

For Porewalker, each structure was uploaded on their webserver and the PDB file containing the marked inner atoms was downloaded. All the residues were extracted from the B-Factor loaded PDB using an in-house python program.

For Moleonline, each structure was submitted to their webserver with the following default parameters. Option to "Ignore HETATOMs" was selected. The probe radius was set to 5 (default) and the interior threshold was taken as 1.1 (default). The origin radius as well as the surface cover radius was taken as 5 and 10 respectively (default). While the weight function was set to the Voronoi Scale (default), the options "Merge Pores" and "Automatic Pores" were turned on. The bottleneck radius, bottleneck tolerance and Max tunnel similarity were taken as 1.2, 3 and 0.7 respectively (default). There were no user-defined starting or ending points specified explicitly. Once the calculations were over, the results were downloaded and the inner residues extracted using an in-house python script.

For CaverWeb, after submitting the structure, the catalytic pocket identified by the tool with the highest pocket score was taken as the starting point for the calculations. In case, CaverWeb was unable to identify a starting point, some inner surface residues were selected manually as indicated in the literature. CaverWeb then converted these into an appropriate starting point for the calculations. All the remaining parameters were set to the default ones. Namely, minimum probe radius was taken as 0.9 while the shell depth and the shell radius were taken as 4 and 3 respectively. The maximal distance was set to 3 with the desired radius set to 5 and the clustering threshold set to 3.5. These default parameters were used for every crystal structure submitted to CaverWeb for the purpose of this study. Upon downloading the results, the inner lying residues were again extracted using a python protocol developed for this purpose.

For CICLOP, all the structures were submitted in the automatic mode (except PDBIDs 1LNQ and 2OAR) of alignment, unless stated otherwise. In cases where conservation scores were calculated, the underlying evolutionary model was selected as 'LG' unless stated otherwise. All the conservation scores calculated for the purpose of this study were done through the empirical bayesian methodology with no exceptions. A list of inner lying residues is provided along with the rest of the results by CICLOP and hence there was no need for a specialized protocol to calculate the same.

Calculation of Time Taken by Each Tool to Provide the Results

The time provided by CICLOP in the results summary file was taken as the measure of CICLOP's runtime. For Porewalker, the time of submission of each crystal structure was recorded and the runtime was calculated as the difference between this recorded time and the receiving time of the email containing the results. Moleonline and Caverweb were timed manually using a stopwatch since they do not provide any other feasible approach to do the same. For timing HOLE 2.2, a standalone version was downloaded and run. Each runtime was calculated using the "time" command in LINUX. Furthermore, the runtime for each method was normalized with respect to CICLOP for each protein structure in the dataset. These values were then summarized as a table showcasing the fold speedup offered by CICLOP.

Molecular Dynamics Simulation for Quantitative Comparison

All atom MD simulations were performed using GROMACS[16;17], version 2020. OPLS-AA/L all atom force field[18], was used to describe the system. SPCE water model[19] was used to solvate the protein and sodium ions were added to neutralize the system. All the simulations were performed at 298K using the modified Berendsen thermostat[20] for temperature control and Verlet cutoff scheme for searching the neighbouring grid cells. Pressure coupling (coupling time 2.0 ps, isothermal compressibility 4.5e-5) using the Parrinello-Rahman scheme was also used in all the simulations due to which the lateral and the perpendicular pressures were coupled independently to maintain a constant pressure of 1 bar. The simulations were performed under periodic boundary conditions in all cardinal directions with the long range electrostatic interactions being treated with the Particle Mesh Ewald (PME) method using a grid-spacing of 0.16 nm combined with fourth-order B-Spline interpolation to compute the potential and forces in between grid points. LINCS algorithm was used to constrain all the bonds. The short range interactions were cut-off at 1.0 nm. The production run for each system was performed for a total of 10 ns with 2 fs being the time step used for numerical integration of the equations of motion. All the starting structures were subjected to a minimization protocol of 50000 steps using the steepest descent algorithm followed by equilibration runs in NVT and then NPT ensembles for 100 ps each.

Each frame post-stabilisation of the protein backbone thus generated by the MD protocol described above was analysed separately using in-house python protocols. Initial frames were discarded in-order to account for short fluctuations in the beginning of the simulation. First all the water atoms lying inside the protein cavity were identified. The distances for each protein atom from the water group identified previously were calculated using GROMACS trjorder. All the residues having at least 1 atom in close proximity (less than or equal to 3.5 Å) to the water group lying inside the cavity were identified as inner residues for the frame in question. Finally, only those residues were considered to be truly inside the cavity which appeared to be lying on the inner surface in at least 90% of the frames. The final list of residues thus generated was taken to be as the "True Positive" set of the inside lying residues, while the rest were categorised as the "True Negatives" for the experiment. Using the residue lists curated in the aforementioned step as true negatives and positives, accuracy and precision was calculated for CICLOP, Porewalker, MoleOnline and CaverWeb.

Analysis of ATP Synthase Structures Using CICLOP

Alignment of the Structures

The bovine ATP synthase F1 domain cavity consisting of $\alpha_3\beta_3$ chains was isolated from the complete crystal structure using UCSF chimera for each of the 7 different sub-states available as crystal structures. The cavity for state 1A was aligned with the Z axis manually and the cavities for the rest of the 6 states were superimposed on the manually aligned structure of state 1A to keep the human error to the minimum. All the 7 cavities were then submitted to CICLOP keeping the "Mode of Alignment" as "Manual".

Plotting Cavity and Gamma Subunit Radius Profiles

The radii at different points along the z axis for the cavity were obtained from CICLOP. For the γ subunit, the radii were estimated using an in-house python protocol. 1 Å thick slices were made along the γ subunit. For each slice, the area of the convex hull of all the atoms in the slice was calculated. A circle having the same area as that calculated by convex hull was constructed and its diameter was used as the final value for that particular slice.

Annotation of the Size of the Cavity

For the annotation of the cavity, chains A and F (that lie opposite to each other) constituting the cavity of the F1 domain of bovine ATP synthase, state 1A, (PDBID: 5ARA) were isolated. The distances to opposite atoms on these chains were calculated and marked using the distance module of UCSF chimera. The corresponding

points on the diameter profile generated by CICLOP were also annotated. For the verification of the length of the cavity, the atoms of chain B having the minimum and the maximum z coordinate were marked. Similarly, the atom having the minimum as well as the maximum Z coordinate for the γ subunit was marked. The distance between these atoms was calculated using the distance module of UCSF chimera. To measure the width of the cavity, the two atoms were manually selected using UCSF chimera at an approximate location where the γ subunit appeared to end.

The Volume and Hydrophobicity Profiles

The total pore volume and total pore hydrophobicities as indicated by CICLOP's result summary file were used to plot the bar graphs using python's matplotlib module.

Evaluating the Orientation of γ subunit During the Rotary Cycle

In order to evaluate the orientation of the γ subunit with respect to the cavity ($\alpha_3\beta_3$), each structure was aligned with respect to chain A. Each constituent chain of the structure was then colored as indicated in the figure. Similarly, to evaluate the orientation of the cavity with respect to γ, each structure was aligned with respect to the γ subunit (chain G). Each constituent chain of the structure was then colored as indicated in the figure using UCSF chimera.

Calculation of Minimum Distance of γ subunit from Each Chain Constituting the Cavity

The γ subunit along with the cavity was isolated from the complete crystal structure for each state and aligned with the Z axis. Moving from the bottom to the top along the z axis, 1 Å thick slices were cut. Any slice which did not contain at least 1 atom from both the cavity (all its constituent chains) and the γ subunit was ignored. In each slice, the distances of all the γ atoms were calculated to all the atoms corresponding to chains A, B, C, D, E and F. The minimum value obtained for each combination was recorded. A similar calculation was performed for each slice and was plotted.

Calculation of the Conservation Scores and Construction of the Conservation Profiles

Calculation of the conservation scores was set to "YES", with the Evalue for the BLAST search as 10 (default), while submitting the structures to CICLOP. The Underlying Evolutionary Model was taken as "REV" with the rate being inferred using an Empirical Bayesian method. The inner residue conservation score marked PDB file containing conservation scores as B factor values was colored according to the scale indicated in the figure. The conservation profile obtained from CICLOP's summary file was used as is.

Calculation of the Dynamic as well as the Constant Residues During State Transitions

Residues that were common in all the seven substates were calculated from the list of residues lining the inner cavity provided by CICLOP. An in house python script was used to calculate the intersection of the seven sets so formed (each set constituting the list of residues detected by CICLOP for that substate). The list of residues so obtained were then marked in pink on each of the structures and visualized using UCSF chimera.

To calculate the residues that change during each state transition, similarly, the residue list provided by CICLOP was used. Considering each list of residues as a set, the difference between two sets (for states that appear one after the other) was calculated using an in house python script. These dynamic residues (calculated for each transition) representing residues that surround the cavity in a particular state but were not present in the previous state, were marked in pink on that structure respectively. All the structures similarly marked were then visualized using UCSF chimera.

References

[1] Marialuisa Pellegrini-Calace, Tim Maiwald, and Janet M Thornton. Porewalker: a novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PLoS Comput Biol*, 5(7):e1000440, 2009.

[2] Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

[3] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12): 2577–2637, 1983.

[4] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[5] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[6] Tal Pupko, Rachel E Bell, Itay Mayrose, Fabian Glaser, and Nir Ben-Tal. Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18(suppl_1):S71–S77, 2002.

[7] Haim Ashkenazy, Shiran Abadi, Eric Martz, Ofer Chay, Itay Mayrose, Tal Pupko, and Nir Ben-Tal. Consurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research*, 44(W1):W344–W350, 2016.

[8] Itay Mayrose, Dan Graur, Nir Ben-Tal, and Tal Pupko. Comparison of site-specific rate-inference methods for protein sequences: empirical bayesian methods are superior. *Molecular biology and evolution*, 21(9): 1781–1791, 2004.

[9] Supratim Choudhuri. *Bioinformatics for beginners: genes, genomes, molecular evolution, databases and analytical tools*. Elsevier, 2014.

[10] Ziheng Yang, Rasmus Nielsen, and Masami Hasegawa. Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular biology and evolution*, 15(12):1600–1611, 1998.

[11] Simon Whelan and Nick Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18(5):691–699, 2001.

[12] Si Quang Le and Olivier Gascuel. An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25(7):1307–1320, 2008.

[13] Jun Adachi, Peter J Waddell, William Martin, and Masami Hasegawa. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast dna. *Journal of molecular evolution*, 50(4):348–358, 2000.

[14] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[15] Thomas D Goddard, Conrad C Huang, Elaine C Meng, Eric F Pettersen, Gregory S Couch, John H Morris, and Thomas E Ferrin. Ucsf chimerax: Meeting modern challenges in visualization and analysis. *Protein Science*, 27(1):14–25, 2018.

[16] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[17] Lindahl, Abraham, Hess, and van der Spoel. Gromacs 2020.4 manual, October 2020. URL `https://doi.org/10.5281/zenodo.4054996`.

[18] George A Kaminski, Richard A Friesner, Julian Tirado-Rives, and William L Jorgensen. Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B*, 105(28):6474–6487, 2001.

[19] Pekka Mark and Lennart Nilsson. Structure and dynamics of the tip3p, spc, and spc/e water models at 298 k. *The Journal of Physical Chemistry A*, 105(43):9954–9960, 2001.

[20] Herman JC Berendsen, JPM van Postma, Wilfred F van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.