MDavocado: analysis and visualisation of protein motion by time-dependent angular diagrams

Boris Gomaz[1] Alessandro Pandini[2] Aleksandra Maršavelski[3] Zoran Štefanić[1]

[1]Division of physical chemistry, Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia

[2]Department of Computer Science, Brunel University London, Uxbridge UB8 3PH, United Kingdom

[3]Department of Chemistry, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000 Zagreb, Croatia

Correspondence: Zoran Štefanić, Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia. Email: zoran.stefanic@irb.hr

Keywords: molecular dynamics, Ramachandran plot, visualization, conformational changes, protein dynamics.

Abstract

Extracting meaningful information from atomistic molecular dynamics (MD) simulations of proteins remains a challenging task due to the high-dimensionality and complexity of data. MD simulations yield trajectories that contain the positions of thousands of atoms in millions of steps. Gaining a comprehensive understanding of local dynamical events across the entire trajectory is often difficult. Here we present a novel approach to visualise MD trajectories in the form of time-dependent Ramachandran plots. Specialised data aggregation techniques are employed to address the challenge of plotting millions of data points on a single image, thereby ensuring that the analysis is independent of the molecule size and/or length of the MD simulation. This approach facilitates quick identification of flexible and dynamic regions, and its strength is the ability to simultaneously observe movements of all amino acids over time. The Python program Mdavocado is freely available at Github (https://github.com/zoranstefanic/MDavocado).

Introduction

While high-resolution techniques such as X-ray crystallography offer detailed insight into biological macromolecules, their major limitation lies in their static nature. By providing a single snapshot structure, they offer limited information on the dynamic behaviour of these macromolecules over time. In reality, the vast majority of biological macromolecules are dynamic entities which employ motions of their constituent parts to carry out functions in living environments. In addition to understanding the structure itself, following their dynamics over time is essential for understanding the mechanisms underlying their actions and interactions. Molecular Dynamics (MD) simulations are now routinely used to elucidate the dynamics behaviour of macromolecules, a fundamental aspect of their function. \cite{Lindahl2015} By estimating the changes in positions of each atom at extremely short time intervals, MD simulations provide a bewildering amount of information. On the other hand, this abundance of data presents the challenge of distinguishing significant and functionally relevant aspects of macromolecular dynamics from random, uncorrelated movement. Due to overwhelming amounts of data generated, some form of data reduction becomes necessary. It then becomes crucial to "wisely" select a suitable data reduction scheme that can effectively capture essential motions, while keeping the process computationally feasible. With access to more powerful computational resources, researchers can run MD simulations of larger and larger molecules on timescales approaching experimental ranges. While this progress is advantageous for obtaining more realistic simulations of molecular motions, the development of efficient methods for representing and visualising the results of these MD simulations lags behind.

To visualise MD simulations, programs such as VMD, \cite{HUMPHREY199633} YASARA \cite{Krieger2014-fd} or Chimera \cite{Pettersen2004-oz} have traditionally been used. These programs are well-known for their ability to visualise MD trajectories and offer extensive analysis capabilities. In the response to the need for porting visualisation capabilities directly to web browsers, without requiring local software installation, several programs have emerged: MDsrv, \cite{Tiemann2017-hi} HTMoL, \cite{Carrillo-Tripp2018-oc} NGLview, \cite{Nguyen_2017} for visualising trajectories within web interface. While these web-based programs may not offer the complete range of capabilities as their desktop counterparts, they do provide the advantage of being available "out-of-the-box" and on the web. As a culmination of these advancements, web-based programs like Mol* Viewer, \cite{Sehnal_2021} have emerged, offering users an unprecedented interactive experience. Such innovations have been adopted by the Protein Data Bank, \cite{Berman_2003} a major database housing all experimentally determined protein structures.

One additional benefit of integrating visualisation directly into the web browser is the ability to combine it with a plethora of already existing tools and technologies for data analysis, yielding very powerful new assets. For example, it is now feasible to visualise protein structures and trajectories with NGLview and combine the analysis with any data analysis library already available in Python using Jupyter notebooks. \cite{Kluyver2016jupyter} This has led to some very interesting analysis tools that can be used for analysing non-covalent contacts throughout the MD trajectories, such as Protein Contact Atlas \cite{Kayikci2018-ob} and GetContacts. \cite{GetContacts} The increasing prevalence of the Python programming language \cite{10.5555/1593511} within the biochemical community resulted in the development of entire packages dedicated to facilitate the analysis of MD simulations, such as pytraj \cite{Roe2013-jx,Amber-MD} and MDAnalysis. \cite{Gowers2016-sx,Michaud-Agrawal2011-do} MDAnalysis is a specialised tool for analysing MD trajectories that provides object oriented representations of data from molecular dynamics trajectories. This library supports direct integration into other programs, expanding the scope of geometrical and energetic analyses. For example, it provides a straightforward method for computing dihedral angles throughout trajectories, such as Ramachandran angles. \cite{Ramachandran1963-un}

The Ramachandran plot has traditionally been used to show the overall distribution of protein main-chain conformations in a single protein structure, representing each amino acid as a single point in a $\phi-\psi$ diagram. This visualisation method readily highlights unusual conformations by identifying points that fall outside the expected ranges of $\phi-\psi$ values, making such diagrams indispensable in protein structure validation. Due to their rather complicated nature and typically large number of data points involved, Ramachandran plots have rarely been used as a tool to follow protein conformation changes over time. Doing so would require overlaying multiple plots, leading to visual clutter and diminishing usability. Recently, there has been a resurgence of interest in visualising dynamic data, \cite{Kozic2024-az} particularly in exploring Ramachandran plots within a dynamic context. \cite{Park2023-po,Rosenberg2023-es,Tam2024-jg}

Time dependent angular diagrams

In this paper, we present a novel approach to the Ramachandran plot that addresses visualisation challenges and demonstrates its potential in tracking protein dynamics. Rather than relying on a single Ramachandran plot, which represents the protein's conformation at a specific time point, we divide the Ramachandran plot into $m-2$ (where m is the number of amino acids) $\phi-\psi$ plots. Each of these plots corresponds to the time evolution of $\phi-\psi$ angles for a single amino acid residue (Figure 1). The first and the last amino acids in the sequence are not included in the analysis as they do not have defined values for both ϕ and ψ angles. The whole MD simulation can be viewed as a sequence of N snapshots, covering a total duration of T. The total simulation time is then divided into n equal segments, each of duration $\Delta t = \frac{T}{n}$. Therefore, each segment will contain $\frac{N}{n}$ data points. As N is orders of magnitude larger than *n*, each plot will still contain a substantial number of data points. These data points accurately capture the movement of each individual amino acid within that specific time segment. We hereafter refer to these as MDavocado diagrams.

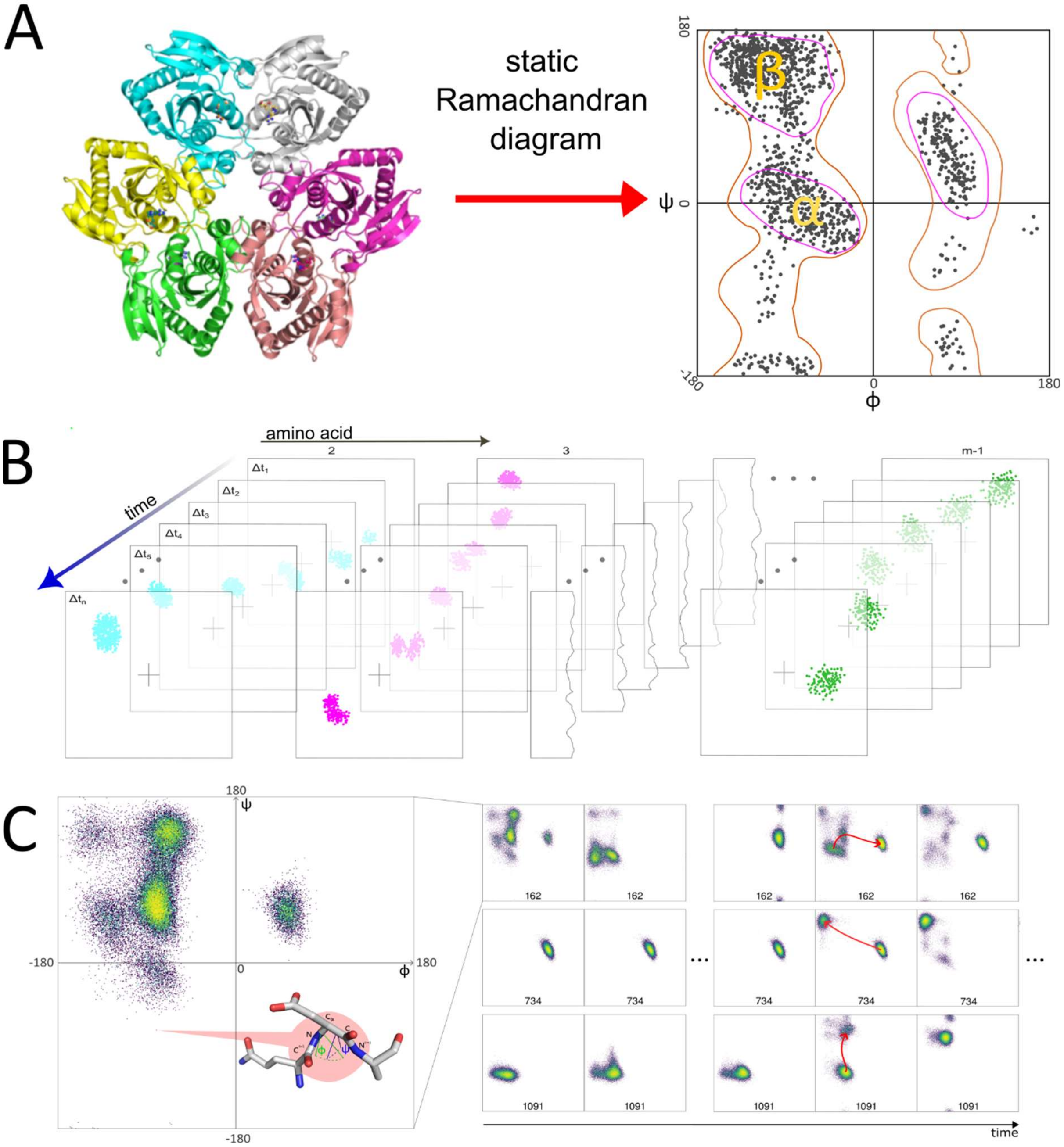

Figure 1. The procedure for constructing MDavocado diagrams. (A) The conformation of a protein at a specific time point can be concisely captured in the Ramachandran diagram, where a static structure is represented as a scatter plot in the $\phi-\psi$ plane, with each amino acid shown as a single dot. This diagram provides an overview of all amino acids simultaneously, illustrating the distribution of amino acids in predominant secondary structure elements like α-helices or β-sheets. (B) To follow the dynamics of the protein over time,

MDavocado diagrams can be used. The total duration of the MD simulation is divided into *n* equal time intervals Δt, and for each amino acid (numbered from 2 to $m-1$) a diagram displaying its $\phi-\psi$ angles is constructed (each snapshot corresponds to a single square). This approach preserves a wealth of information of the movement of each amino acid in time. These individual snapshots can then be aggregated over time. (C) Each dot on the plot represents a specific combination of ϕ and ψ angles for a particular amino acid at a specific moment in time (as shown in the inset, indicated by pink arrow). Due to potential overlap of points, an aggregation technique is used to visualize the density of points using a colour scheme (in this case from dark blue over green to yellow, indicating increasing density). The overall figure is produced as a collection of all $\phi-\psi$ pairs within one segment of the trajectory Δt. A series of such plots is generated for each amino acid, which may sometimes occupy different regions, indicating conformational changes (as highlighted by the red arrows). The numbers represent the ordinal position of the amino acids in primary protein sequence.

However, while reducing Ramachandran plots to individual amino acids simplifies the representation, a challenge arises when plotting the diagrams for a single time interval due to the overwhelming number of points involved. To address this issue, we employed a technique for data reduction and visualisation implemented in the Python library Datashader. \cite{datashader} This technique allows the visualisation of an arbitrarily large number of points on a two-dimensional plane, ensuring that there are no overlaps. In essence, it projects the points onto the plane, aggregates the results and represents the number of points (point density) in each pixel using a suitable colour coding scheme. Furthermore, the technique automatically scales the aggregation process to optimise visibility and dynamic range, ensuring that all data remains visible.

As a result of the data reduction procedure, a series of images is generated that accurately represent the conformational changes of each amino acid throughout the entire MD simulation (Figure 1C). To further emphasise movement of amino acids in time, these images can be sequentially arranged in time to illustrate their movement throughout the trajectory (Figure S1), producing an MDavocado diagram for each amino acid. Additionally, these images can be combined into larger panels to show chains, domains or even entire protein structure (an example of MDavocado animated diagram is given as Figure S2). This approach offers the advantage of providing a comprehensive overview of the entire MD simulation in a single dynamic figure or movie. This visualisation method gives an immediate overview of which local regions of the protein are rigid, and which are more flexible, showing changes in position over time. Moreover, the synchronised nature of these changes in time can reveal temporal relationship and potential causal connection between different local motions.

Case studies

To assess the utility of MDavocado diagrams in extracting crucial conformational information from MD simulation data, we have applied this method in two distinct protein case studies. To observe potential structural changes, the proteins were subjected to MD simulations of 1 microsecond total duration. These selected proteins vary significantly in size and are recognized for their distinct conformational changes. The application of this analytical tool to both case studies confirmed the simplicity, accuracy and effectiveness of the workflow.

Case study 1: Acyl carrier protein (ACP)

We have applied our newly developed tool to a small-size protein, ACP, which consists of only 77 amino acids, and it is a well-known and well-described system. \cite{Gally1978-wn,Worsham2003-jk} ACP plays a central role in the fatty acid synthesis and exists in three forms: apo-ACP, holo-ACP, and acyl-ACP. Apo-ACP is an inactive acyl carrier protein without the prosthetic group, holo-ACP has the attached phosphopantetheine prosthetic group that serves as a carrier of a growing acyl chain, while acyl-ACP refers to the protein with a covalently attached growing acyl chain to the prosthetic group. Here, we performed a molecular dynamics (MD) simulation of apo-ACP and decanoyl-ACP, which exhibits conformational change following the transition of the decanoyl chain from being solvent-exposed to being sequestered into the hydrophobic pocket of ACP.

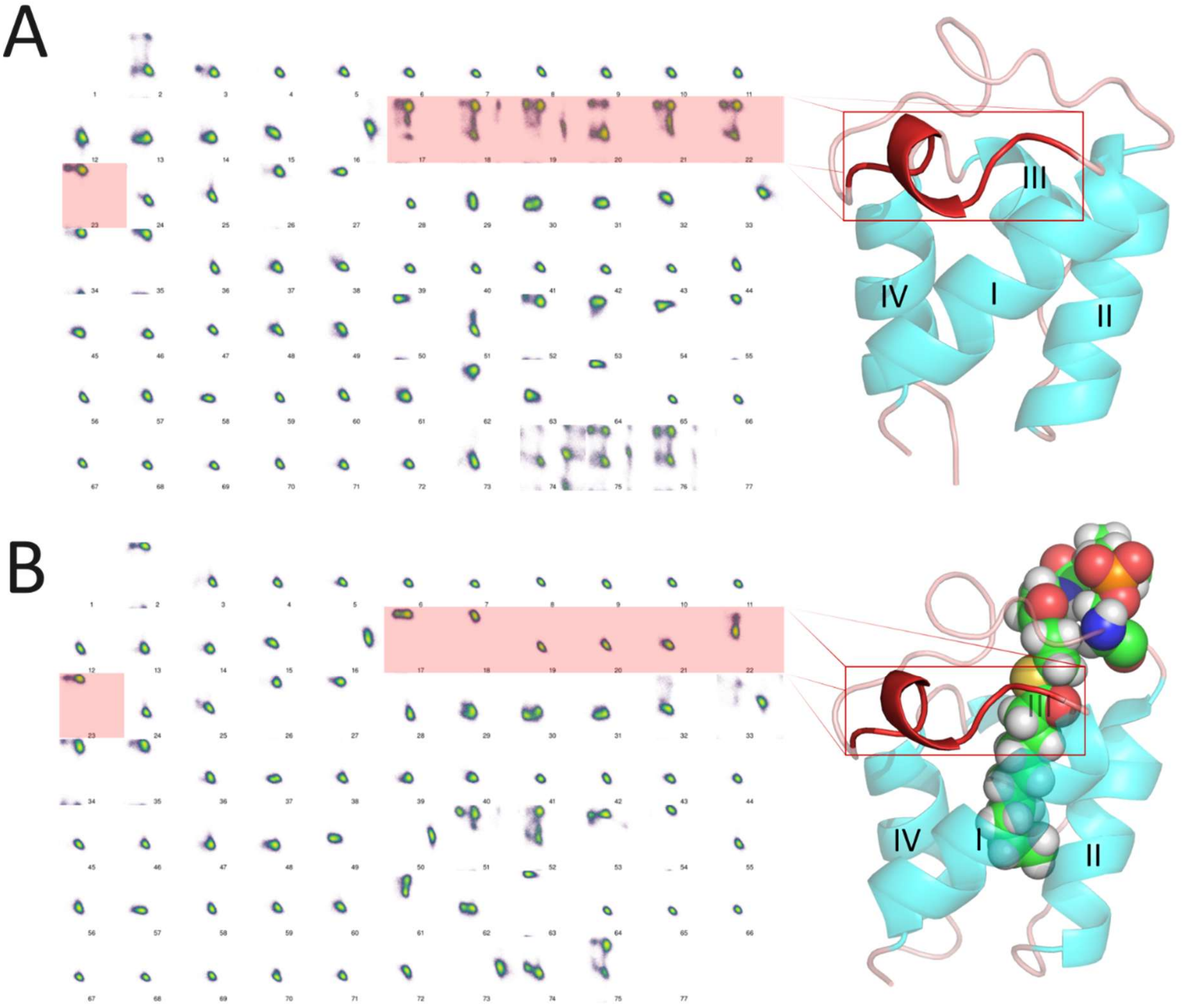

Figure 2. Structures and MDavocado diagrams of ACP. (a) In the apo-ACP simulation, highlighted in red are highly flexible region of the protein (from Gly17 to Val23). The corresponding part of MDavocado diagram (connected

with thin lines) shows significant motion (Figure S3). (b) In the decanoyl-ACP simulation, the binding of the prosthetic group's into the protein structure leads to the stabilisation of the red region. This stabilisation is clearly visible in the MDavocado diagram, indicating stability in that specific region (Figure S4). On the MDavocado diagram it is easy to spot the characteristic signature of short stable α-helix formed by the amino acids 19-21.

Previous observations have shown that simulations of \textit{E. coli} apo-acyl carrier protein (ACP) exhibit regions with elevated root mean square fluctuations (RMSF) values. Specifically, the longest loop I (Val18 to Asp36), which connects helices I and II, demonstrates high flexibility. \cite{Chan2008-ea} However, upon acylation, the decanoyl-ACP enters the hydrophobic pocket on the side of the protein between helix II and loop II. The entrance of the decanoyl chain stabilises the protein, especially the backbone of loop I in the region from Gly17 to Val23, which adopts a short helix structure as evident from our analysis and movies derived from MD simulations (Supplementary MovieS1.mov and MovieS2.mov). This stabilisation occurs in the absence of direct contact between the sequestered decanoyl chain and amino acids in the region Gly17 to Val23 of loop I (Figure 2).

Case study 2: Purine nucleoside phosphorylase (PNP)

The second protein of interest was PNP, a large oligomeric enzyme that has 233 amino acids in each of its six monomeric units, making it approximately 18 times bigger than ACP. PNP plays a crucial role in the purine salvage pathway, which is one of the two pathways for synthesising purines. Some bacteria, such as \textit{Helicobacter pylori}, lack enzymes in the \textit{de novo} pathway, and therefore rely on the purine salvage pathway, making PNP essential for their survival. \cite{Roszczenko-Jasinska2020-kw} Given that \textit{H. pylori} is an invasive bacteria, PNP becomes an interesting target for the development of new antibiotics. \cite{Bubic2023-mk} Understanding the dynamics of PNP is crucial for this purpose. Crystallographic structures revealed two distinct conformations of the active site of this enzyme: open and closed. Depending on whether substrates are bound in the active site or not, these structures can undergo a conformational change of the terminal α-helix H8, the last of the eight helices in the structure. This helix undergoes a segmentation around Ser220 resulting in the formation of two smaller helices and the closure of the active site. \cite{Narczyk2018-wg} Here, we performed a molecular dynamics (MD) simulation of one hexameric structure composed of 4 monomeric units in closed and 2 monomeric units in open conformation.

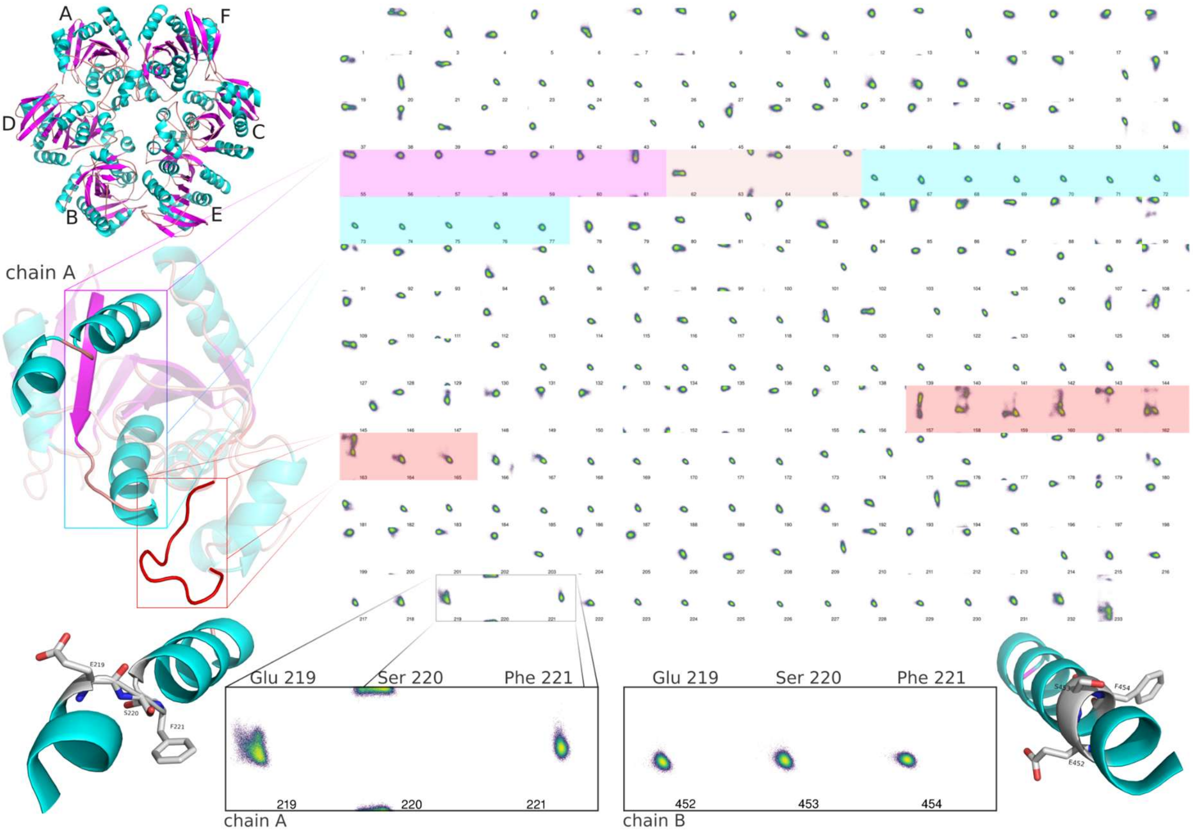

Figure 3. The structure of purine nucleoside phosphorylase from \textit{H. pylori} (full hexamer shown, PDB code 6F4X, top left) and the full MDavocado diagram of the chain A on the right (number 1 is missing due to undefined $\phi$ angle for the first amino acid; a dynamic version is available as Figure S2). The MDavocado plots show remarkable richness and diversity of conformations, and distinct signatures of $\beta$-sheets, loops and $\alpha$-helices are clearly visible. As an illustration, on the left, chain A is enlarged and one central $\beta$-loop-$\alpha$ region (amino acids 55-77) is indicated on the corresponding part of the MDavocado diagram. It clearly shows characteristic elements of secondary structure; also a highly mobile loop region from residues 157-165 is highlighted in red. Interestingly, this loop is situated at the interface between the A and D chains, which form a functional dimer, and not on the protein surface. The terminal helices H8 in chain A (bottom left) and B (bottom right) are shown. The three amino acids affected by segmentation (219-221) and their corresponding parts in MDavocado diagrams are shown for both chain A (indicated by lines) and chain B. A remarkable difference between the two chains is evident, marked by a kink in the -helical pattern at amino acid 220.

What is immediately noticeable from MDavocado diagrams produced from PNP MD simulations is the remarkable similarity among all six chains (all six chains from one particular MD simulation can be seen at \urlq{https://alokomp.irb.hr/md/trajectory/1458}). This is hardly surprising because the secondary structure is of course highly conserved. It is also worth noting that some parts of the protein remain very stable, with minimal changes, while other parts exhibit more variability. The striking observation is that changes in amino acid conformations occur abruptly, often transitioning by sudden jumps between several stationary states. These stationary states are clearly visible on the MDavocado diagrams as distinct and separated blobs in $\phi-\psi$ plane. This indicates that the characteristic duration of these distinct states is comparable to the total simulation time, which is readily apparent from observing the visual representation of the entire MD simulation trajectory. Representing the time dimension in the form of a movie offers the advantage of capturing simultaneity between transitions. Namely, it allows for the identification of distinct parts of the protein undergoing coordinated changes very close in time, even if these parts are not necessarily close together in physical space. This observation may be the signature of intrinsic correlation between different regions of the protein, indicating potential communication pathways or allosteric routes. Additionally, subtle differences between different parts of the protein can also be observed, such as clear distinction between different conformational states. For example, the one can easily differentiate between open and closed conformation of chains in PNP enzyme just by viewing certain small parts of MDavocado diagram for those chains (Figure 3).

Conclusions

We have presented a novel approach for tracking the dynamics and local conformational changes of proteins. By employing molecular dynamics simulations and visual representations, we have effectively captured the intricate movements and structural alterations occurring within the protein over time. The utilisation of a time-resolved approach, presented in the form of a movie, provides a unique perspective on simultaneous changes occurring across different regions of the protein, unveiling potential communication pathways and allosteric routes that may govern its function. The methodology showcased in this research not only enables the observation of coordinated changes in distant regions of the protein but also highlights the subtle differences that exist between various parts of the molecule, offering a comprehensive understanding of its local dynamics. The identification of abrupt transitions between distinct conformational states further emphasises the complexity and versatility of the protein's structural dynamics. Moving forward, this approach holds significant promise for advancing our understanding of biomolecular systems. The integration of computational simulations with experimental data can further enhance the accuracy and predictive power of this approach, paving the way for innovative discoveries in the field of structural biology and drug design.

AUTHOR CONTRIBUTIONS

Boris Gomaz: Conceptualization (equal); investigation (equal); methodology (equal); software (equal); visualisation (equal); writing – review and editing (equal).

Alessandro Pandini: Investigation (equal); software (supporting); writing – review and editing (equal).

Aleksandra Maršavelski: Investigation (equal); methodology (equal); software (supporting); writing – review and editing (equal).

Zoran Štefanić: Conceptualization (equal); funding acquisition (lead); investigation (equal); methodology (equal); project administration (equal); software (equal); supervision (equal); visualisation (equal); writing – original draft (lead); writing – review and editing (equal).

Data availability statement

The source code for MDavocado is freely available on the Github page: \url{https://github.com/zoranstefanic/MDavocado}. The program was used extensively as part of the project ALOKOMP (\url{https://alokomp.irb.hr/}) for visualization and analysis of the MD simulations of PNP enzymes, where many more examples of PNP trajectories can be explored.

Acknowledgements

Computational resources for the MD simulations were provided by the Isabella cluster and Supek supercomputer at the University Computing Centre (SRCE). This research was funded by the Croatian Science Foundation (grant IP-2019-04-6764), CSF Mobility program - outgoing mobility of assistants (project grant MOBDOK-2023-6223) and the Development Research Support (NPOO.C3.2.R2-I1.06).