Bayesian statistical method for detecting structural and topological diversity in polymorphic proteins

Shuto Hayashi 1,*, Jun Koseki 1, Teppei Shimamura *
Division of Systems Biology, Graduate School of Medicine, Nagoya University, Nagoya, Japan

A R T I C L E   I N F O

Article history:
Received 15 August 2022
Received in revised form 17 November 2022
Accepted 18 November 2022
Available online 21 November 2022

Keywords:
Protein structural diversity
Molecular dynamics simulation
Persistent homology
Topological data analysis
Bayesian statistical model

A B S T R A C T

Polymorphisms in immune-related proteins and viral spike proteins are high and complicate host-virus interactions. Therefore, diversity analysis of such protein structures is essential to understand the mechanism of the immune system. However, experimental methods, including X-ray crystallography, nuclear magnetic resonance, and cryo-electron microscopy, have several problems: (i) they are conducted under different conditions from the actual cellular environment, (ii) they are laborious, time-consuming, and expensive, and (iii) they do not provide information on the thermodynamic behaviors. In this paper, we propose a computational method to solve these problems by using MD simulations, persistent homology, and a Bayesian statistical model. We apply our method to eight types of HLA-DR complexes to evaluate the structural diversity. The results show that our method can correctly discriminate the intrinsic structural variations caused by amino acid mutations from the random fluctuations caused by thermal vibrations. In the end, we discuss the applicability of our method in combination with existing deep learning-based methods for protein structure analysis.

1. Introduction

Proteins are macromolecular compounds consisting of chains of amino acids. The amino acid sequence defines the three-dimensional structure and physical properties of the protein, and a slight difference in the amino acid sequence can alter its activity and its binding specificity and affinity to the target molecule. In particular, some immune-related proteins, such as immunoglobulins, T-cell receptors, and major histocompatibility complexes (MHCs), are highly polymorphic within the same species and even within the same individual. On the other hand, viral spike proteins, which interact with cell surface receptors on host cells, are also polymorphic and evolve to evade the host immune mechanism and increase the infectivity. Thus, in order to understand the immune mechanism, it is necessary to clarify how these polymorphisms lead to differences in protein structures, intramolecular and intermolecular interactions, and phenotypes.

To this end, experimental methods, including X-ray crystallography, nuclear magnetic resonance, and cryo-electron microscopy, have been widely used to solve protein structures at atomic resolution. However, there are several problems with using these experimental methods to analyze protein structural diversity. First, the experiments are conducted under conditions that are highly different from the actual cellular environment. Therefore, there is no guarantee that the protein has the same structure in a cell. Second, it is infeasible to apply the experimental methods comprehensively to various mutant proteins because they are laborious, time-consuming, and expensive. Finally, the protein structure obtained by the experimental methods is only a snapshot and does not provide information on its thermodynamic behavior in a cell.

To overcome these problems, molecular dynamics (MD) simulation can be used to obtain multiple protein conformations. In MD simulation, the trajectory of all the atoms is calculated by numerically solving Newton's equations of motion. Since initial mutant protein structures for MD simulation can be created by substituting amino acid residues using a single template protein, there is no need to experimentally determine the structures for all the mutant proteins. Although MD simulation has greatly advanced structural biology [1–13], experience and expertise are required to analyze, understand, and interpret the results of MD simulation. Recently, persistent homology, a method used in topological data analysis, has been used in various types of research fields to capture the topological features of datasets [14–21]. Persistent
homology has the advantage of being able to quantitatively evaluate the structural information of datasets without experience or expertise. However, since previous methods use persistent homology to characterize entire protein structures, it is difficult to find differences in local substructures of mutant proteins. In fact, mutant proteins do not necessarily differ in their entire structures, but often only in their local substructures.

In this paper, we propose an in-silico approach that combines persistent homology with MD simulation to extract the topological features of proteins and detects the intrinsic variations in diverse protein substructures using a Bayesian statistical model. We applied our method to several types of HLA-DR molecules, which are one of the human MHC class II molecules, to evaluate the structural diversity. We demonstrate that our method can correctly discriminate the intrinsic structural variations caused by amino acid mutations from the random fluctuations by thermal vibrations.

2. Material and methods

First, three-dimensional structures of control and case proteins are created from a PDB file of a template protein. Then, MD simulations are performed to sample the protein structures that fluctuate due to thermal vibrations. Persistent homology is further applied to the sampled protein structures to capture the topological features. Lastly, differences in the topological features between the control and case proteins are scored using a Bayesian statistical model. The details are described in the following subsections.

2.1. Preparation of control and case protein structures

Three-dimensional structures of control and case proteins are needed to be created as inputs to MD simulations. First, a PDB file of a template protein is downloaded from the Protein Data Bank [22]. The template protein structure is then used to create the control and case protein structures by substituting amino acid residues with the Schrödinger’s Maestro’s “mutate” option [23]. For each of the control and case structures, Schrödinger’s Protein Preparation Wizard [24,25] is used to cap N- and C-termini with acetyl and N-methyl groups, create disulfide bonds between close cysteine residues, and add hydrogens and optimize hydrogen bonding networks.

2.2. Structure sampling by MD simulation

MD simulation is used to sample the control and case protein structures. In our method, Amber 2018 [26] is used to prepare input files and perform energy minimization and MD simulation. First, a three-dimensional simulation box is created for each protein. Each protein is solvated in a TIP3P water box [27] with a buffer distance of 12 angstroms, in combination with the Amber ff99SB force field [28] and periodic boundary conditions. Na+ and Cl− ions are also added to neutralize the charges of the system. Then, energy minimization is performed to find the stable arrangement of the atoms in the system. The stabilized structure is used as the initial coordinates of MD simulation. The simulation is performed over 100 picoseconds (500,000 time steps with a step size of 0.2 fs) while gradually heating the system from a temperature of 0 K to 310 K. The coordinates of the atoms in the system are saved at every 1,000 steps. All the 500 conformations are used in the subsequent analysis.

2.3. Structural feature extraction by persistent homology

To identify the structural differences between the control and case proteins, structural features are extracted from the coordinates sampled by the MD simulations. Persistent homology can be used for this purpose, which is an approach that captures the topological properties of data points (Fig. 1A). In persistent homology, balls centered at the data points with a common radius are placed, and the radius is gradually increased (Fig. 1B). As the radius is increased, the balls intersect to form loops (1-dimensional cycles) and voids (2-dimensional cycles) (Fig. 1C). As the radius is further increased, the loops and voids disappear (Fig. 1D). These appearance and disappearance times of the cycles, called birth and death times, reflect the topological characteristics of the data points and are used to analyze the shape formed by the data points (Fig. 1E).

First, the coordinates of alpha carbons are extracted from the sampled protein structures to obtain their backbone structures. To focus on important structural variations, alpha carbons in intrinsically disordered regions are identified by the DSSP algorithm [29] and excluded. Persistent homology is then applied to the remaining alpha carbons using the TDA package in R [30], which produces a time series of the birth and death times for each cycle of each protein. Since each cycle is not created for all the conformations, the number of obtained birth and death times could be less than 500. In addition, a cycle observed in the control protein is not necessarily observed in the case protein, and vice versa. Therefore, only cycles that satisfy the following two conditions are used in the subsequent analysis: (i) the cycle is obtained at least five times for both proteins, and (ii) the cycle is obtained within the first five steps for both proteins.

2.4. Detection of structural variations between control and case proteins

The time series of the birth and death times of a given cycle indicates how the atoms that make up the cycle thermally vibrate. To detect the structural variations between the control and case proteins, intrinsic structural variations must be distinguished from thermal vibrations. Furthermore, the thermal vibrations might not be fully captured when the number of obtained birth and death times is insufficient. To address these problems, we propose a method to identify intrinsic structural variations based on a Bayesian statistical model.

For each cycle, let \( \{b_1^{(1)}, d_1^{(1)}\} \) and \( \{b_2^{(2)}, d_2^{(2)}\} \) be the birth and death times at time \( t \) for the control and case proteins. For simplicity, the subscript for the cycle is omitted. To reduce the correlation that the lifespan \( b_1^{(1)} - d_1^{(1)} \) and \( b_2^{(2)} - d_2^{(2)} \) are used instead of the death times. Let \( \mathbf{x}_1 = (b_1^{(1)}, d_1^{(1)})^T \) and \( \mathbf{x}_2 = (b_2^{(2)}, d_2^{(2)})^T \), then the observations are generated from the following hierarchical Bayesian statistical model: For each \( p \in \{1, 2\} \),

\[
\Sigma^{(p)} \sim W^{-1}(\Psi_0, \nu_0),
\]

\[
\mu^{(p)} | \Sigma^{(p)} \sim \mathcal{N}(\mathbf{m}_0, \Sigma^{(p)} / \kappa_0),
\]

\[
\mathbf{x}^{(p)} \sim \mathcal{N}(\mu^{(p)}, \Sigma^{(p)}),
\]

where \( W^{-1} \) is an inverse Wishart distribution, \( \mathcal{N} \) is a normal distribution, and \( \Psi_0 \in \mathbb{R}^{2 \times 2}, \nu_0 \in \mathbb{R}, \mathbf{m}_0 \in \mathbb{R}^2, \) and \( \kappa_0 \in \mathbb{R} \) are constant parameters that control the prior distributions. Here, we set \( \Psi_0 \) to the identity matrix and \( \nu_0 \) to 2 so that the inverse Wishart distribution is a non-informative prior. Similarly, we set \( \kappa_0 \) to \( 0 \) so that the normal distribution is also a non-informative prior. In this case, the normal distribution becomes an improper prior, but
it is guaranteed to have a proper posterior because the birth and death times are obtained at least five times. The values of \( m_0 \) do not affect the posterior distribution because \( \kappa_0 \) is +0.

Since the prior distributions are conjugate, the posterior distributions of \( \mu^{(p)} \) and \( \Sigma^{(p)} \) can be analytically calculated as follows:

\[
\Sigma^{(p)} \mid \left\{ x^{(p)}_t \right\}_t \sim \mathcal{N}^{-1}(\Psi^{(p)}, \psi^{(p)}),
\]

\[
\mu^{(p)} \mid \Sigma^{(p)}, \left\{ x^{(p)}_t \right\}_t \sim \mathcal{N} \left( m^{(p)}, \Sigma^{(p)} \right),
\]

where

\[
\Psi^{(p)} = \Psi_0 + S^{(p)},
\]

\[
\psi^{(p)} = v_0 + N^{(p)},
\]

\[
m^{(p)} = \bar{x}^{(p)},
\]

\[
\kappa^{(p)} = N^{(p)},
\]

and

\[
N^{(p)} = \left| \left\{ x^{(p)}_t \right\}_t \right|,
\]

\[
\bar{x}^{(p)} = \frac{1}{N^{(p)}} \sum_{t} x^{(p)}_t,
\]

\[
S^{(p)} = \sum_{t} \left( x^{(p)}_t - \bar{x}^{(p)} \right) \left( x^{(p)}_t - \bar{x}^{(p)} \right)^T.
\]

The posterior distributions of \( \mu^{(p)} \) and \( \Sigma^{(p)} \) represent the extent to which the thermal vibrations cause the cycle to fluctuate in the protein.

To detect the structural variations between the control and case proteins, \( \mu^{(1)} \)'s, \( \Sigma^{(1)} \)'s, \( \mu^{(2)} \)'s, and \( \Sigma^{(2)} \)'s are sampled from the posterior distributions and compared. If the cycle observed in the control protein is similar in structure to the case protein, the distributions for the control protein \( \mathcal{N}(\mu^{(1)}, \Sigma^{(1)}) \) and the case protein \( \mathcal{N}(\mu^{(2)}, \Sigma^{(2)}) \) should be almost the same. Conversely, if the cycle observed in the control protein has an intrinsically different structure from the case protein, \( \mathcal{N}(\mu^{(1)}, \Sigma^{(1)}) \) and \( \mathcal{N}(\mu^{(2)}, \Sigma^{(2)}) \) would be dissimilar. To measure the similarity of the distributions, the Jeffreys divergence [31] is used. The Jeffreys divergence of two probability distributions \( p \) and \( q \) is defined by

\[
D_{\text{J}}(p, q) = D_{\text{KL}}(p || q) + D_{\text{KL}}(q || p),
\]

where \( D_{\text{KL}} \) is the Kullback-Leibler divergence [32] and given by

\[
D_{\text{KL}}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx.
\]

The Jeffreys divergence is larger than or equal to 0, and similar two distributions have a small Jeffreys divergence. The Jeffreys divergence of two normal distributions can be analytically calculated because the Kullback-Leibler divergence of two normal distributions is also analytically given by

\[
D_{\text{KL}}(\mathcal{N}(\mu_1, \Sigma_1) || \mathcal{N}(\mu_2, \Sigma_2)) = \frac{1}{2} \left( \log \frac{\Sigma_2}{\Sigma_1} - d + \text{tr} \left( \Sigma_2^{-1} \Sigma_1 \right) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \right),
\]

where \( d \) is the dimension of the normal distributions. Thus, the Jeffreys divergences can be calculated from each of the sampled \( \mu^{(1)} \)'s, \( \Sigma^{(1)} \)'s, \( \mu^{(2)} \)'s, and \( \Sigma^{(2)} \)'s.

Since the parameters are sampled multiple times, the Jeffreys divergences can also be calculated multiple times. The structural variation in the cycle between the control and case proteins is detected using a score defined by the 5th percentile of the calculated Jeffreys divergences, which we call a structural variation score (SV score). If the SV score is larger than a certain threshold \( D^* \), the Jeffreys divergences are determined to be significantly larger than 0. This means that the distributions of the birth and death times are different between the control and case proteins, and that these two proteins have different structures in the cycle. In our method, we set \( D^* \) to 4.

### 3. Results

#### 3.1. Application of our method to HLA-DR complexes

We applied our method to HLA-DR molecules to demonstrate its capability. First, we downloaded an X-ray crystal structure of the HLA-DRA1*01:01 and HLA-DRB5*01:01 complex (PDB ID: 1H15) to be used as a template. By replacing amino acid residues in the template beta chain, we created eight mutant HLA-DR complexes (DRB1*04:05, DRB1*08:02, DRB1*08:03, DRB1*12:01, DRB1*13:02, DRB1*14:54, DRB1*15:01, and DRB1*15:02), which are common in Japanese. We performed MD simulation and...
persistent homology for these complexes as described in Material and Methods to obtain cycles. **Fig. 2A** shows the number of cycles observed in each complex, and **Fig. 2B** shows how many cycles each residue formed, which indicates that several regions tend to form cycles across the complexes. We then applied our Bayesian statistical model to each combination of the complexes. **Fig. 2C** shows the number of cycles observed in both complexes of each combination, and **Fig. 2D** shows the number of cycles in which structural variations were detected by our model. Below, we will show several examples of structural variations captured by our method.

### 3.2. Loop variation detected across HLA-DR complexes

First, we show a loop in which the structural variation was detected across the complexes. **Fig. 3A** shows the scatter plots of the birth and death times of the loop. This figure implies that the complexes can be classified based on the topological similarity of the loop: cluster 1 (the DRB1*04:05, DRB1*12:01, and DRB1*13:02 complexes) and cluster 2 (the DRB1*14:54 and DRB1*15:02 complexes). Although the DRB1*08:02 complex has a similar structure to cluster 1, it has larger birth times. The loop structures in the DRB1*08:02 and DRB1*15:01 complexes fluctuate and move between the two clusters. We calculated the SV scores using the DRB*04:05 complex as control and the other complexes as cases (**Fig. 3B**). The SV scores for the DRB1*08:02, DRB1*14:54, and DRB1*15:02 complexes were higher than the criterion $D' = 4$, while the SV scores for the DRB1*12:01 and DRB1*13:02 complexes were smaller compared to the other SV scores, which means that our scoring method can correctly distinguish topological clusters. We also checked the root-mean-square deviations (RMSDs) between the average loop structure of the DRB1*04:05 complex, which was obtained by extracting and superimposing the loop conformations of the DRB1*04:05 complex and taking the average of the coordinates, and the loop conformations of the eight complexes (**Fig. 3C**). Comparison of **Fig. 3B** and 3C suggests that the SV scores are correlated with the RMSDs.

### 3.3. Loop variation detected between similar complexes

Next, we show the only loop in which the structural variation was detected between the DRB1*15:01 and DRB1*15:02 complexes, which have the same amino acid sequence except that the 86th amino acid residue is a valine and a glycine, respectively (**Fig. 4A**). This loop consists of five amino acid residues in the alpha chain DRA*01:01 (V34, K39, E40, Q57, and A59) located near the 86th amino acid residue of the beta chain. It has been reported that the 86th amino acid residue of the beta chain affects the stability of the HLA-DR complex [33], suggesting that it is reasonable that the mutation in the beta chain induces the structural variation in the alpha chain. Furthermore, because the loop is located at the peptide-binding groove, the structural change could alter the peptide-binding motif. **Fig. 4B** shows the scatter plot of the birth and death times of the loop. We also checked the RMSDs between the average loop structure of the DRB1*15:02 complex and the loop conformations of the two complexes (**Fig. 4C**). Although the loop was observed only 5 and 13 times in the DRB1*15:01 and DRB1*15:02 complexes, respectively, our method was able to detect the structural variation by using the Bayesian statistical model.

---

**Fig. 2.** Statistics of cycles observed in the HLA-DR complexes. (A) The number of cycles observed in each complex. (B) The number of cycles consisting of each amino acid residue in each complex. (C) The number of cycles observed in both complexes of each combination. (D) The number of cycles in which structural variations were detected by our method.
To demonstrate the effectiveness of our method, we checked whether structural differences in DRB1*15:01 and DRB1*15:02 can be detected by RMSDs alone. We compared the average structures between the two complexes and found no significant difference (RMSD = 0.744Å). We also found that the structural fluctuations were small in the two complexes (standard deviations of RMSDs = 0.097Å and 0.063Å for DRB1*15:01 and DRB1*15:02, respectively). Thus, there is no significant difference in the entire structures between DRB1*15:01 and DRB1*15:02, which makes it difficult to find substructural differences in these two complexes using RMSDs alone.

3.4. Void variation detected among several complexes

Finally, we show a void in which the structural variation was detected among the DRB1*04:05, DRB1*08:03, DRB1*12:01, and DRB1*13:02 complexes. Fig. 5A shows the scatter plots of the birth and death times of the void, which suggests that the DRB1*08:03
complex has a different structure from the other complexes. We calculated the SV scores using the DRB*04:05 complex as control and the other complexes as cases (Fig. 5B). As expected, only the SV score for the DRB1*08:03 complex was higher than the criterion $D^* = 4$. We also checked the RMSDs between the average void structure of the DRB1*04:05 complex and the void conformations of the four complexes (Fig. 5C). Since the SV scores are correlated with the RMSDs, it can be said that our method can capture structural variations characterized not only by loops but also voids.

4. Discussion

In this paper, we proposed a method for detecting protein structural variations using MD simulations, persistent homology, and a Bayesian statistical model. Our method has several advantages. First, our method uses multiple protein structures sampled by MD simulation rather than a single snapshot of the protein structure. Therefore, it is possible to determine whether the difference between two protein structures is intrinsic or due to thermal vibrations. Second, our method is quantitative using persistent homology and does not require experience or expertise in protein structure analysis. Third, our method can correctly identify structural variations in cycles, even when they were observed only several times. This is because our method can measure the credibility of the inference result by sampling the parameters and the Jeffreys divergences in our Bayesian statistical model. Finally, our method focuses on differences in the minimal structural components, such as loops and voids, rather than differences in the coordinates of atoms. Hence, our method has the potential to efficiently extract substructural variations that may alter the functions.

Although the threshold for SV scores was set to 4 in this study, it needs to be chosen depending on how different substructures should be detected because SV scores reflect the magnitudes of structural differences rather than statistical significances. Theoretically, if the two covariance matrices of birth–death times are equal between control and case proteins ($\Sigma_1 = \Sigma_2 = \Sigma$), then the SV score is equal to the square of the Mahalanobis' distance between the two means $((\mu_2 - \mu_1)^T \Sigma^{-1} (\mu_2 - \mu_1))$. In this study (SV threshold = 4), this implies that the two means are more than 2 S.D. apart. By using a larger SV threshold, only substructures with large differences can be detected.

Recent advances in deep learning technology have enabled highly accurate biomolecular structural analysis in silico, such as protein structure inference [34–37], molecular conformation generation [38–40], and protein–ligand docking [41–43]. A large amount of structural information on biomolecules can be easily obtained, and structural biology has been greatly influenced. By combining with the technology, our method would help accelerate this research field elucidate the biological meanings of protein diversities.

CRediT authorship contribution statement

Shuto Hayashi: Conceptualization, Methodology, Software, Writing – original draft, Funding acquisition. Jun Koseki: Conceptualization, Formal analysis, Writing – review & editing, Funding acquisition. Teppei Shimamura: Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by grants from KAKENHI grants-in-aid for young scientists (S.H., grant number 22 K17993), KAKENHI grants-in-aid for scientific research (B) (J.K., grant number 22H03686), KAKENHI grants-in-aid for scientific research (B) (T.S., grant number 20H04281), KAKENHI grants-in-aid for...
scientific research on innovative areas on information physics of living matters (T.S., grant number 22H04839), KAKENHI grants-in-aid for challenging exploratory research (T.S., grant number 20K21832) from the Japan Society for the Promotion of Science (JSPS). It was also supported by RADDAR-J (T.S., grant number JP20ek0109488), Brain/MINDS Health and Diseases (T.S., grant number JP21wm0425007) from the Japan Agency for Medical Research and Development (AMED), and Moonshot R&D (T.S., grant number JPMJMS2025) from the Japan Science and Technology Agency (JST).

The computational resources were provided by AI Bridging Cloud Infrastructure (ABCI) at the National Institute of Advanced Industrial Science and Technology (AIST), TSUBAME3.0 at Tokyo Institute of Technology, and SHIROKANE at the Human Genome Center, the University of Tokyo.

References

[1] Dauber X, Jaun B, Seebach D, van Gunsteren WF, Mark AE. Reversible peptide folding in solution by molecular dynamics simulation. J Mol Biol 1998;280:925–32.
[2] Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Nat Struct Biol 2002;9:646–52.
[3] Karplus M, Kuriyan J. Molecular dynamics and protein function. Proc Natl Acad Sci U S A 2005;102:6679–85.
[4] Ho BK, Dill KA. Folding very short peptides using molecular dynamics. PLoS Comput Biol 2006;2:e27.
[5] Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. Atomic-level characterization of the structural dynamics of proteins. Science 2010;330:341–6.
[6] Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. Science 2011;334:517–20.
[7] Raval A, Piana S, Eastwood MP, Dror RO, Shaw DE. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. Proteins 2012;80:2071–9.
[8] Tautermann CS, Seeliger D, Kriegl JM. What can we learn from molecular dynamics simulations for GPCR drug design? Comput Struct Biotechnol J 2015;13:111–21.
[9] Chevalier A, Silva D-A, Rocklin GJ, Hicks DR, Vergara R, Murapa P, et al. Massively parallel de novo protein design for targeted therapeutics. Nature 2017;550:74–170.
[10] Liu X, Shi D, Zhou S, Liu H, Liu H, Yao X. Molecular dynamics simulations and novel drug discovery. Expert Opin Drug Discov 2017;12:23–37.
[11] Latorraca NR, Wang JK, Bauer B, Townshend RJL, Hollingsworth SA, Olivierie JE, et al. Molecular mechanism of GPCR-mediated arrestin activation. Nature 2018;557:452–6.
[12] Pantzar T. The current understanding of KRAS protein structure and dynamics. Comput Struct Biotechnol J 2019;18:189–98.
[13] Casarotto PC, Grych M, Fred SM, Kovalova V, Moliner R, Enkavi G, et al. Antidepressant drugs act by directly binding to TRKB neurotrophin receptors. Cell 2021;184:1299–313.
[14] Petri G, Expert P, Turkheimer F, Carhart-Harris R, Nutt D, Hellyer PJ, et al. Homologous scaffolds of brain functional networks. J R Soc Interface 2014;11:20140873.
[15] Kovacevic-Nikolic V, Bubenik P, Nikolaic D, Heo G. Using persistent homology and dynamical distances to analyze protein binding. Stat Appl Genet Mol Biol 2016;15:19–38.
[16] Hiroaka Y, Nakamura T, Hirata A, Escolar EM, Matsue K, Nishiyama Y. Hierarchical structures of amorphous solids characterized by persistent homology. Proc Natl Acad Sci U S A 2016;113:7035–40.
[17] Pike JA, Khan AO, Pallini C, Thomas SG, Mund M, Ries J, et al. Topological data analysis quantifies biological nano-structure from single molecule localization microscopy. Bioinformatics 2020;36:1614–21.
[18] Townsend J, Micucci CP, Hymel JH, Maroulas V, Vogiatzis KD. Representation of molecular structures with persistent homology for machine learning applications in chemistry. Nat Commun 2020;11:3230.
[19] Chung Y-M, Hu C-S, Lo Y-L, Wu H-T. A Persistent Homology Approach to Heat Rate Variability Analysis With an Application to Sleep-Wake Classification. Front Physiol 2021;12:637684.
[20] Cole A, Loges G, Shiu S. Qualitative and interpretable order parameters for phase transitions from persistent homology. Phys Rev B 2021;104:104426.
[21] Solis JM, Orlova T, Bednarska K, Lesiak P, Wolinski TR, D’Alessandro G, et al. Tracking the time evolution of soft matter systems via topological structural heterogeneity. Commun Mater 2022:3:1.
[22] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.
[23] Maestro, Schrödinger, LLC, New York, NY, 2021.
[24] Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2021; Impact, Schrödinger, LLC, New York, NY; Prime, Schrödinger, LLC, New York, NY, 2021.
[25] Ulisson M, Sandergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J Chem Theory Comput 2011;7:525–37.
[26] D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham III, W.V.D. Cruzeno, et al. AMBER 2018, University of California, San Francisco, 2018.
[27] Jorgensen WL, Chandrasekhar J, Madura JD. Comparison of simple potential functions for simulating liquid water. J Chem Phys 1983;79:926–35.
[28] Hornak V, Abel R, Okura A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 2006;65:712–25.
[29] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1981;22:3577–637.
[30] B.T. Fasy, J. Kim, F. Lecci, C. Maria, D.L. Millman, V. Rouvreau, TDA: Statistical Tools for Topological Data Analysis. R package version 1.8.7. https://CRAN.R-project.org/package=TDA, 2022.
[31] Jeffrey H. Theory of probability. 2nd edition. Oxford: Clarendon Press; 1948.
[32] Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat 1951:22:79–86.
[33] Verreck FA, Termijtelen A, Koning F. HLA-DR beta chain residue 86 controls DR alpha beta dimer stability. Eur J Immunol 1993;23:1346–50.
[34] Junger J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.
[35] Baek M, Dimaio F, Anschichen I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021;373:871–6.
[36] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv 2022. https://doi.org/10.1101/2022.07.20.500998.
[37] Wu R, Ding F, Wang R, Shen R, Zhang X, Luo S, et al. High-resolution de novo structure prediction from primary sequence. bioRxiv 2022. https://doi.org/10.1101/2022.07.31.009999.
[38] Ma XO, Lu Y, Yao S, Liang J, Shi S, Eromn J, Tang G. DeepDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In International Conference on Learning Representations, 2022.
[39] B. Jing, G. Corso, J. Chang, R. Barzilay, T. Jaakkola. Torsional Diffusion for Molecular Conformer Generation. In ICLR2022 Machine Learning for Drug Discovery, 2022.
[40] E. Hoogeboom, V.G. Satorras, C. Vignac, M. Welling. Equivariant Diffusion for Molecule Generation in 3D. In Proceedings of the 39th International Conference on Machine Learning, 162, 8867-8887, 2022.
[41] M.R. Masters, A.H. Mahmoud, Y. Wei, M.A. Lill. Deep Learning Model for Protein Conformer Generation. In ICLR2022 Machine Learning for Drug Discovery, 2022.