اولین همایش ملی ریاضیات زیستی

دانشگاه نیشابور

و أنْ لَیسَ لِلْإِنسانِ إلاّ ما سَعی سوره نجم- آیه ۳۹

و اینکه برای انسان جز حاصل تلاش او نیست

گواهی ارائه مقاله

بدین وسیله گواهی می شود، مقاله با عنوان :

"persistent homology for prediction of protein folding"

توسط: هانیه میرابراهیمی، آمنه بابایی و اعظم بابایی

مورد پذیرش هیأت داوران و کمیته علمی جهت ارائه در **"اولین همایش ملی ریاضیات زیستی"** که ۲۱ اسفند ماه ۱۳۹۷ در دانشگاه نیشابور برگزار شد، قرار گرفته است. این مقاله توسط **هانیه میرابراهیمی** ارائه گردید.

امید است این گواهی در بهبود هر چه بیشتر عملکرد ایشان در راستای افزایش بهره وری و تحقق توسعه پایدار در بخش های مختلف ریاضیات زیستی کشور عزیزمان ایران موثر واقع شده و در ارتقاء علمی ایشان مد نظر قرار گیرد.

دکتر مژگان افخمی گلی
رئیس دانشکده علوم پایه

دکتر سید محسن صالح
دبیر علمی همایش

FIRST ANNUAL NATIONAL CONFERENCE ON BIOMATHEMATICS
March 12-13, 2019

Persistent homology for prediction of protein folding

Hanieh Mirebrahimi[1]

Department of pure mathematics, Ferdowsi university of Mashhad, Mashhad, Iran

Ameneh Babaee

Azam Babaee

Department of biophysics, Tarbiat Modares university of Tehran, Tehran, Iran

Abstract

Topological data analysis in an approach to use topological techniques for analysis of datasets. Persistent homology is one of the main tools of topological data analysis used for reducing the dimension and complexity of the data sets, and also to distinguish topological features and delete noises. Site directed mutagenesis is widely used to understand the structure and function of biomolecules. Computational prediction of protein mutation impacts offers a fast, economical and potentially accurate alternative to laboratory mutagenesis.

In this talk, topology based mutation predictor (T-MP) is introduced to dramatically reduce the geometric complexity and number of degrees of freedom of proteins, while element specific persistent homology is proposed to retain essential biological information. The present approach is found to outperform other existing methods in globular protein mutation impact predictions.

Keywords: persistent homology, protein folding, topological data analysis

AMS Mathematical Subject Classification [2010]: 55U10, 55U05

1 Persistent homology

Persistent homology is an algebraical tool to investigate topological features. In topology, a group structure was defined called homology group for simplicial complexes, denoted by H_i indexed by $i = 0, 1, 2, \ldots$ related to the dimension of holes. We can simplify homology to the number of its generators denoted by β_i, called ith Betti number, counting the connected components for $i = 0$, cycles for $i = 1$, holes for $i = 2$, voids for $i = 3$ and ... in a simplicial complex. By a simplicial complex, we mean a set of points called vertices, a subset of pair of vertices called edges, a subset of triple of vertices that are pairly joined by edges, called triangles, and ....

Consider a set of points in the euclidean space (see Figure 1). Start with a distance α, and connect pairs of points that are no further apart than α. Then, we fill in complete simplexes, that is if we find three points connected by edges that form a triangle, we fill in the triangle with a 2-dimensional face. Any 4 points that

[1] speaker

are all pairwise connected get filled in a 3-simplex and .... The resulting simplicial complex is called the Rips complex or the Vietoris-Rips complex. We then apply homology to this complex which reveals the presence of the holes.

Note that each hole appears at some particular value of α, namely α_1 and disappears at another α, namely α_2. We can represent the persistence or age of this hole as a pair (α_1, α_2). See Figure 1.

A collection of such bars is called a barcode, and barcodes are essential objects of study in persistent homology. See Figure 1.

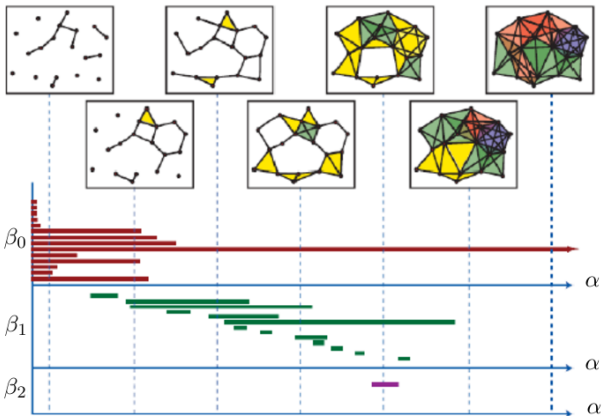

Figure 1: Horizontal axis shows the distance and vertical axis shows the number of holes, called Betti number.

The small holes, having short persistence, are not important in our data analysis and represented by short bars in the barcodes, but larger holes, having long persistence, can be considered as significant feature of the data, and they are represented by a long bar in the barcode. Therefore, the brief interpretation of the barcodes can be stated as: short bars represent noise, and long bars represent features.

A key property of barcodes is that they are stable under perturbations of the data. In other words, if you move a point a little bit, the barcode only changes a little bit. This stability of barcodes is caused by topological invariance of the homology groups, and it is important in applications in dynamical cases. For more details and applications see [2].

Standard algorithms exist to compute barcodes such as streaming algorithm. Streaming algorithms are algorithms for processing data streams in which the input is presented as a sequence of items and can be examined in only a few passes.

The worst case runtime of these algorithms is cubed in the number of simplices, although the complication is really worst case. In addition, improving data structures by topological simplification can speed up the computation significantly.

2 Protein stability upon mutation

Mutagenesis, as a basic biological process that changes the genetic information of organisms, serves as a primary source for many kinds of cancer and heritable diseases, as well as a driving force for natural evolution. In laboratories, site directed mutagenesis analysis is a vital experimental procedure for exploring protein functional changes. Nonetheless, site directed mutagenesis analysis is both time-consuming and expensive. Computational prediction of protein mutation impacts is an important alternative to experimental

mutagenesis analysis for the systematical exploration of protein structural instabilities, functions, disease connections, and organism evolution pathways. A major advantage of these approaches is that they provide an economical, fast, and potentially accurate alternative to site directed mutagenesis experiments.

The last class of approaches is knowledge based methods that invoke modern machine learning techniques to uncover hidden relationships between protein stability and protein structure as well as sequence. A major advantage of knowledge based mutation predictors is their ability to handle increasingly large and diverse mutation data sets.

A common challenge for all existing mutation impact prediction models is in achieving accurate and reliable predictions of membrane protein stability changes upon mutation.

A key feature of all existing structure based mutation impact predictors is that they either fully or partially rely on direct geometric descriptions which rest in excessively high dimensional spaces resulting in large number of degrees of freedom. In practice, the geometry can easily be over simplified. Mathematically, topology, in contrast to geometry, concerns the connectivity of different components in a space, and offers the ultimate level of abstraction of data. However, conventional topology incurs too much reduction of geometric information to be practically useful in biomolecular analysis. Persistent homology, a new branch of algebraic topology, retains partial geometric information in topological description, and thus bridges the gap between geometry and topology. It has been applied to biomolecular characterization, identification and analysis. However, conventional persistent homology makes no distinction of different atoms in a biomolecule, which results in a heavy loss of biological information and limits its performance in protein classification. In the present work, we introduce element specific persistent homology (ESPH), interactive persistent homology and binned barcode representation to retain essential biological information in the topological simplification of biological complexity. We further integrate ESPH and machine learning to analyze and predict protein mutation impacts. The essential idea of our topological mutation predictor (T-MP) is to use ESPH to transform the biomolecular data in the high-dimensional space with full biological complexity to a space of fewer dimensions and simplified biological complexity, and to use machine learning to deal with massive and diverse data sets. A distinct feature of the present T-MP is that the prediction results can be analyzed and interpreted in physical terms to shed light on the molecular mechanism of protein folding energy changes upon mutation. Additionally, the mathematical model for different types of mutations can be adaptively optimized according to the performance analysis of ESPH features.

3 Persistent homology characterization of protein

Unlike physics based models which describe protein folding in terms of covalent bonds, hydrogen bonds, electrostatic and van der Waals interactions, the natural language of persistent homology is topological invariants, i.e., the intrinsic features of the underlying topological space.

When persistent homology is used to analyze three dimensional (3D) protein structures, one-dimensional (1D) persistent homology barcodes are obtained as topological fingerprints (TFs).

As an illustration, we consider the persistent homology analysis of a wild type protein (PDB:1ey0) and its mutant. The mutation (G88W) occurred at residue 88 from Gly to Typ is shown at Figure 2 a and b. In this case, a small residue (Gly) is replaced by a large one (Typ). We carry out persistent homology analysis of a set of heavy atoms within 6A from the mutation site. Persistent homology barcodes of the wild type and the mutant are respectively given in Figure 2 c and d.

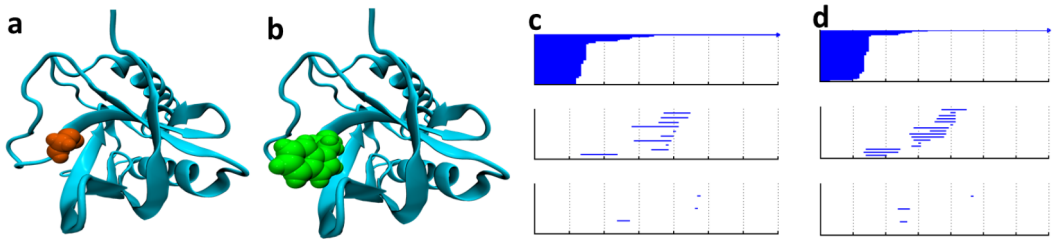

Figure 2: An illustration of persistent homology barcode changes from wild type to mutant proteins [1].

The above topological representation of proteins does not contain sufficient biological information, such as bond length distribution of a given type of atoms, hydrogen bonds, hydrophobic and hydrophilic effects, to offer an accurate model for protein mutation impact predictions. To characterize chemical and biological properties of biomolecules, we introduce element specific persistent homology (ESPH). Instead of labeling every atom as in many physics based methods, we distinguish different element types of biomolecules in constructing persistent homology barcodes. For proteins, commonly occurring element types include C;N;O; S and H.

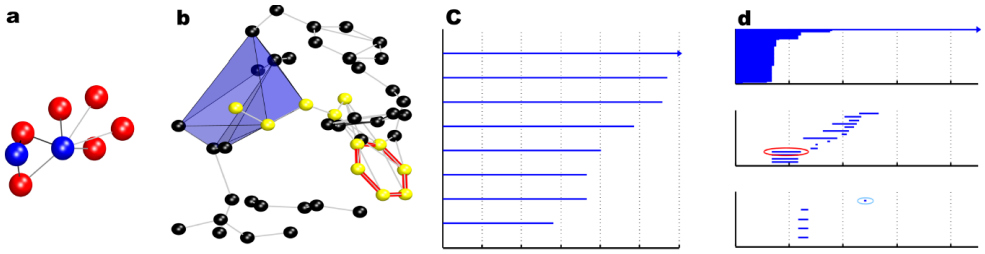

Figure 3: An illustration of element specific persistent homology (ESPH) indicating the hydrophilic network (Left) and hydrophobic network (Right) at a mutation site [1].

The most important issue in protein mutation impact analysis is the interactions between the mutation site and the rest of the protein. To describe these interactions, we propose interactive persistent homology adopting the distance function $DI(A_i; A_j)$ describing the distance between two atoms A_i and A_j defined as

$$DI(A_i; A_j) = \begin{cases} \infty; \text{if} Loc(A_i) = Loc(A_j); \\ DE(A_i; A_j); \text{otherwise}; \end{cases}$$

where $DE(\cdot, \cdot)$ is the Euclidean distance between the two atoms and $Loc(\cdot)$ denotes the location of an atom which is either in a mutation site or in the rest of the protein. In the persistent homology computation, Vietoris- Rips complex (VC) and alpha complex (AC) are used for characterizing first order interactions and higher order patterns respectively. To characterize interactions of different kinds, we construct persistent homology barcodes on the atom sets by selecting one certain type of atoms in mutation site and one other certain type of atoms in the rest of the protein.

Barcodes computed by persistent homology are capable of revealing the molecular mechanism of protein stability. For example, interactive ESPH barcodes generated from carbon atoms are associated with hydrophobic interaction networks in proteins. Similarly, interactive ESPH barcodes between nitrogen and oxygen atoms correlate to hydrophilic interactions and/or hydrogen bonds as shown in Figure 3. Interactive ESPH barcodes are also able to reveal other bond information; notwithstanding, they can not always be interpreted as covalent bond, hydrogen bonds, or van der Waals bonds in general. In fact, interactive ESPH

barcodes provide an entirely new representation of molecular interactions.

While the topological descriptors give a through examination of the atomic arrangements and interactions, some other crucial properties are not explicitly characterized. Additionally, due to the diverse quality of the structures examined, some higher level descriptors such as residue level descriptors can enhance the robustness of the model. Therefore, we include some auxiliary descriptors from the aspect of geometry, electrostatics, amino acid types composition, and amino acid sequence. The geometric descriptors contain surface area and van der Waals interaction. The electrostatics descriptors are consisted of atomic partial charge, Coulomb interaction, and atomic electrostatic solvation energy. The high level descriptors include neighborhood amino acid composition and predicted pKa shifts. The sequence descriptors describe the secondary structure and residue conservation score collected from Position-specific scoring matrix.

The topological features and the auxiliary features are ideally suited for being used as machine learning features to predict protein stability changes upon mutation. We have examined a number of machine learning algorithms, including decision tree learning, random forest, and gradient boosted regression trees (GBRTs).

To demonstrate the power of the proposed T-MP for protein mutation impact predictions, we consider a data set of 2648 mutation instances in 131 proteins, called S2648 data set.9 Additionally, a subset of the S2648 data set involving 67 proteins, named S350 set, is used as a test set.

A comparison of the performances of various methods is summarized in Table 1. Pearson correlations coefficient (RP) and RMSE for test set S350, and five-fold cross validations for training set S2648, are given for various methods, including ours.

A comparison between T-MP-1 and T-MP-2 indicates that geometric, electrostatic and sequence features give rise to approximately 5% improvement over the original topological prediction, indicating the importance of geometric, electrostatic and sequence information to mutation predictions.

Our topology based approaches significantly outperform other existing physical or empirical methods. When auxiliary features are used together with topological features, a 5% improvement in Pearson correlation coefficient is found. Compared with Rosetta- MP, which achieves the best performance with terms designed for membrane proteins,19 the present T-MP-2 has a 84% higher Pearson correlation coefficient. Nonetheless, Kroncke et als statement about membrane protein mutation impact predictions still holds as the best Pearson correlation coefficient is only 0.57 and the best RMSE is over 1 kcal/mol. We therefore call for further methodology developments to improve membrane protein mutation impact predictions.

This article introduces element specific persistent homology to appropriately simplify biomolecular complexity while effectively retain essential biological information in protein mutation impact predictions. Extensive numerical experiments indicate that element specific persistent homology offers some of the most efficient descriptions of protein mutation impacts that cannot be obtained by other conventional techniques.

References

[1] Z. Cang and G. Wei wei, Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. Bioinformatics, 33(22), 3549–3557, 2017.

[2] A. Zomorodian, Topology for Computing, Cambridge University Press, Cambridge, 2005.

e-mail: h.mirebrahimi@um.ac.ir
e-mail: am.babaee@mail.um.ac.ir
e-mail: aa.babaee2020@gmail.com