Higher School of Economics - National Research University

Faculty of Computer Science

Data Science

Internship Report

Student	Zhaoyu Guo
Place of Internship (HSE department/Full name of a company, contacts)	HSE University, Faculty of Computer Science, International Laboratory of Algebraic Topology and Its Applications
Assignment	Development of the persistent Tor-algebra approach to data analysis
Internship Supervisor	Ivan Limonchenko, Associate Professor

Internship Report

Zhaoyu Guo

September 2022

1 Introduction

Protein-protein interactions, or PPIs, are crucial to the interactomics system in all living cells and are involved in the majority of biochemical processes. This paper proposed persistent Tor-algebra (PTA), through the combination of Tor-algebra and the filtration process, PTA-based molecular characterization and featurization, and PTA-based stacking ensemble learning model for PPIs. Topological data analysis (TDA) is used in molecular representation and featurization.

2 Persistent Tor-algebra

Assume we have filtration of the simplicial complex:

$$\emptyset = \mathcal{K}_0 \to \mathcal{K}_1 \to \dots \mathcal{K}_n = \mathcal{K} \tag{1}$$

It is a sequence of nested simplicial complexes connected by inclusions, where $\mathcal{K}_i$ is a subcomplex of $\mathcal{K}_{i+1}$, so we can construct its face ring $\mathbb{F}(\mathcal{K}_i)$.

Firstly, let $f : k[\mathcal{K}_1] \to k[\mathcal{K}_2]$ be an isomorphism of K-algebras, and we can assume f is a graded isomorphism and by restrictions to the linear components, so there are $\mathcal{K}_1$ and $\mathcal{K}_2$ be two simplicial complexes on the vertex sets $[m_1]$ and $[m_2]$, $[m_1] = [m_2]$, so F: $k[m_1] \to k[m_2]$.

$$k[v_1, \ldots, v_{m_1}] \xrightarrow{F} k[v_1, \ldots, v_{m_2}] \tag{2}$$

$$k[\mathcal{K}_1] \xrightarrow{f} k[\mathcal{K}_2] \tag{3}$$

We will have the isomorphism $f^* : X(\mathcal{K}_2) \to X(\mathcal{K}_1)$, which is the restriction of the K-linear isomorphism $F^* : k^{m_2} \to k^{m_1}$, so the isomorphism f^* establishes a bijective correspondence

$$\phi : (\text{maximal faces of } \mathcal{K}_2) \to (\text{maximal faces of } \mathcal{K}_1) \tag{4}$$

which is defined by the formula $f^*(S_I) = S_{\phi(I)}$, where I is a maximal face of $\mathcal{K}_2$. So for $i_1, i_2 \in$[m$_1$], we put $i_1 \quad i_2$ if and only if the two sets of maximal faces $\mathcal{K}_1$ containing i_1 and i_2 respectively coincide. The equivalence classes in $[m_1]$ are the minimal nonempty intersections of maximal faces of $\mathcal{K}_1$, and similarly for $[m_2]$. And there are the same numbers of elements in the corresponding equivalence classes. So there exists a bijective map $\varphi : [m_2] \to [m_1]$. So the face ring detemines its underlying simplicial complex.

And we can derive a series of face rings:

$$\mathbb{F}(\mathcal{K}_0) \to \mathbb{F}(\mathcal{K}_1) \to \cdots \to \mathbb{F}(\mathcal{K}_n) \tag{5}$$

where two adjacent face rings are connected by the homomorphism induced from the inclusion map. We can construct the minimal resolution(R_{min}, d) of the K[m]-module $k[\mathcal{K}]$. Then $R^0_{min} \cong 1 \cdot k[m]$ is a free module with one generator of degree 0. And the basis of R^{-1}_{min} is a minimal generator set for I_k, which corresponds to the missing faces of K. Then map d: R^{-1}_{min} $\rightarrow$R$^0_{min}$. Thus minimal resolution can be obtained, its Tor-algebra Tor($\mathcal{K}_i$) can be constructed and the homomorphism naturally induces homomorphism between the Tor-algebra of two adjacent face rings. We get a sequence of Tor-algebra connected by homomorphisms

$$Tor(\mathcal{K}_0) \rightarrow Tor(\mathcal{K}_1) \rightarrow \cdots \rightarrow Tor(\mathcal{K}_n) \tag{6}$$

More, by considering a specific (i, 2j)-th graded component of Tor-algebra modules, we get

$$Tor^{-i,2j}(\mathcal{K}_0) \rightarrow Tor^{-i,2j}(\mathcal{K}_1) \rightarrow \cdots \rightarrow Tor^{-i,2j}(\mathcal{K}_n) \tag{7}$$

This is called the persistent Tor-algebra of K in the (i, 2j)-th graded component.

3 Method

This paper presents a persistent Tor-algebra based stacking model, containing two basic learners the 1D convolutional neural network and gradient boosting tree, and two meta learners. and from the decomposition, the persistent Tor-algebra barcode is exactly the direct sum of the persistent coho-

mology barcode of the subcomplexes. The 144 features are divided into two types of learners, the first with 72 Tor-algebra type features as a 1D CNN model, in addition, auxiliary features are combined with topological features to form 72 gradient boosting tree models. Construct a Vietoris-Rips complex K. For each atomic combination, compute the persistent Tor-algebra components.

Experimental results show that PTA-based stacking ensemble learning (PTA-SEL) models perform better than other traditional models. Testing in the AB-Bind database and SKEMPI database, charged, polar, special cases except for hydrophobic and alanine mutations, and non-alanine all remain consistent. This model has achieved the best results with PCC of 0.866 and RMSE of 1.216 kcal/l when performed on the SKEMPI S1131 dataset. And performance on the AB-Bind S645 dataset where in different mutation regions except for the support regions is consistent. The topological algebraic method gives a quantitative characterization of the molecular structure, and this is the first time proposed that persistent Tor-algebra is used for the molecular representation and featurization combined with machine learning.

[1] [2]

References

[1] Victor M Buchstaber and Taras E Pano. *Toric topology*, volume 204. American Mathematical Soc., 2015.

[2] LIU Xiang and Kelin Xia. Persistent tor-algebra based stacking ensemble learning (pta-sel) for protein-protein binding affinity prediction. In *ICLR*

2022 Workshop on Geometrical and Topological Representation Learning, 2022.