 doi:10.20944/preprints202204.0027.v1

A Network Embedding Approach for Annotating Protein Structures.

Barbara Puccio[1], Ugo Lo Moio,,, Luisa Di Paola[3], Pietro Hiram Guzzi[1,2[0000−0001−5542−2997]], and Pierangelo Veltri[1]

1 , Department of Surgical and Medical Sciences, University of Catanzaro
{barbara.puccio,hguzzi,veltri}@unicz.it,
2 IEEE SENIOR MEMBER, ACM Professional Member
, 3 Unit of Chemical-Physics Fundamentals in Chemical Engineering, Department of Engineering, University Campus Bio-Medico di 27 Roma, via Ãlvaro del Portillo 21, 00128 Rome, Italy,

Abstract. Protein Contact Network (PCN) is an emerging paradigm for modelling protein structure. A common approach to interpreting such data is through network-based analyses. It has been shown that clustering analysis may discover allostery in PCN. Nevertheless Network Embedding has shown good performances in discovering hidden communities and structures in network. In this work, we compare some approaches for graph embedding with respect to some classical clustering approaches for annotating protein structures.

1 Introduction

Proteins are polymers made of twenty different amino acids organised to assemble a linear chain. The linear sequence of the amino acid determines the spatial conformation of proteins. The spatial structure of proteins is characterized by the presence of a central carbon atom (called carbon-C), a carboxyl group, an amino group and a lateral chain, different for each amino acid (this chain can be hydrophobic, no polar or charged). They are linked to each other by covalent bonds (that are called peptide bonds) between molecules.

The amino acids sequence is known as primary structure. The secondary structure indicates the folding of peptides, i.e. protein subsquences, chain resulting from the interaction between each amino acid and neighboring amino acids. The main types of secondary structure are α-helix and β-sheet. The tertiary structure is a combination of secondary structures, that makes a complex molecular shape (3D-shape). A protein in its 3D-conformation is called 'native' and this is closely connected with its biological function. Finally, the quaternary structure is only present in proteins with multiple subunits (peptide chains) that can be the same or different.

In such a scenario Protein Contact Networks (PCNs) emerged as a relevant paradigm for the analysis of protein molecular structures [7]. A PCN is a graph built from protein structure. A node in a PCN represents an Carbon-α atom of

the backbone, while an edge represents a spatial distance of the atoms in the range 4 and 8 Å.

PCN descriptors are useful to model and analyse protein functions [7]. PCNs allow to identify modules in protein molecules through network spectral clustering [11,**?**], with relevant application in different biological contexts ([**?**,**?**].

For instance, analysis of PCNs allows to detect such as allosteric regulation. *Allostery* is the ability of proteins to trasmit a signal from one site to another in response to environmental stimuli and this is related to the trasmission of information across the protein from a sensor site (or allosteric site) to the effector site (or binding site)[9]. Allostery may also be studied using wet lab methods such as X-ray or NMR structures correspondent to different activation states or molecular dynamics simulation of allosteric agent binding. Such methods are usually time and resource consuming, therefore there is the need to introduce computational method to detect allostery and then to annotate allosteric regions.

Starting from PCNs it is possibile to detect modules in protein structure using clustering algorithm approach. In a graph a cluster is a group of nodes that are characterized by a strong intra-cluster connection (in terms of number of contacts) and a weaker inter-cluster connection. Clustering allows to detect community (cluster) in a graph and this is perfectly comparable to the modules detection in protein structure. Two methods have been devised to partition PCNs into clusters: a geometrical method, based on the k-means algorithm and spectral clustering, and the clustering of the embedding of the network.

PCNs allows to simplify protein analysis to detect modules, essential in allosteric regulation. On the other hand graphs consist of high number of nodes and links (particularly in protein world) thus it is challenging to apply different mathematical and statistical operations. In this situation, embeddings appear as a reasonable solution. Based on potential of graph embeddings, we propose a PCNs analysis using clustering approaches on embeddings, in order to discover allostery in PCN and annotate protein structures.

We use PCN-Miner, a software tool implemented in the Python programming language able to import protein in the Protein Data Bank format and generate the corresponding protein contact network. Also it offers a set of algorithms for the PCN analysis[5]. As previously reported, in this work we focus on application of clustering algorithm on the embeddings with the aim of evaluating network embedding approaches in PCN analysis. Our analysis based on SARS-CoV-2 spike glycoprotein, in its closed form, and some of its Variants of Concern.

2 Protein Contact Networks

A protein structure can be represented as a complex three-dimensional object, formally defined by the coordinates in 3D space of its atoms. Despite the large availability of protein molecular structures data, there are yet many problems regarding the relationship between protein structures and their functions. For this reason it is necessary to define simple descriptors that can describe protein structures with few numerical variables. Structure and function are based on the

 doi:10.20944/preprints202204.0027.v1

complex network of inter-residue interactions, where residues are identified by amino acids sequences [3]. Therefore, the residues interactions are used to define protein interaction networks. Protein interaction networks are thus used to study protein functions. The most simple choice to define networks, is to represent the protein structure by means of α-carbon location. The spatial position of C_α is still reminiscent of the protein backbone and this allows to highlight also the most important characteristics of the three-dimensional structure. Starting from the C_α spatial distribution, a distance matrix d is evaluated where each $d_{i,j}$ represents the Euclidean distance in the 3D space between the i-th and j-th residues, defined as

$$d_{i,j} = \sqrt{((x_i - x_j)^2) + ((y_i - y_j)^2) + (z_i - z_j)^2)} \tag{1}$$

where (x_i, y_i, z_i) and (x_j, y_j, z_j) respectively are the cartesian coordinates of residue i and j. Matrix d is used to define a Protein Contact Network concept, that is an alternative and different representation of using graph-based models to represent protein structures. A graph is the most natural structure to represent proteins, where nodes (or vertices) are the protein residues and links (or edges) between the i-th and the j-th nodes (residues) represent residue contacts. In the graph represetation there exists a link between two residues i and j if the distance between two residues (i.e., $d_{i,j}$) is higher than 4 and lower than 8 Å. The lower end excludes all covalent bonds, which are not sensible to environment change (so to protein functionality), while the upper end gets rid of weaker non-covalent bonds (so not significant for protein functionality). At this point, it is possible to build up adjacency matrix A, whose generic element is defined as:

$$A_{ij} = \begin{cases} 1 & \text{if} \quad 4\mathring{A} \leq d_{ij} \leq 8\mathring{A} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The adjacency matrix of a graph is unique in regard to the ordering nodes. In the case of proteins, in which the order of nodes (residues) corresponds to the residues sequence (primary structure) it can be said that its corresponding network is unique: this establishes a one-to-one correspondence between protein and its network.

3 Protein Contact Network Analysis

Graph embeddings are the transformation of property graphs to a vector or a set of vectors. Embedding should capture the graph topology, vertex-to-vertex relationship, and other relevant information about graphs, subgraphs, and vertices. There are a few reasons why graph embeddings are needed:

1. Graphs consist of edges and nodes. Those network relationships can only use a specific subset of mathematics, statistics, and machine learning, while vector spaces have a richer toolset of approaches.

2. Adjacency matrix describes connections between nodes in the graph. It is a | V | x | V | matrix, where | V | is a number of nodes in the graph. Each column and each row in the matrix present a node. Non-zero values in the matrix indicate that two nodes are connected. Using an adjacency matrix as a feature space for large graphs is almost impossible. Imagine a graph with 1M nodes and an adjacency matrix of 1M x 1M. Embeddings are more practical than the adjacency matrix since they pack node properties in a vector with a smaller dimension.
3. Vector operations are simpler and faster than comparable operations on graphs.

The development of novel methods fro encoding structural information of graph to be used for subsequent analysis is a recent area in research. These methods are usually referred to as *graph representation learning* or *graph-embedding* [6]. The goal of these approaches is learning a mapping for graph substructures (i.e. nodes or sub graphs) into points of a low-dimensional vector space R^d, having $d < n$, n is the dimension of the adjacency matrix [6]. We here focus on node-based embeddings, thus all these methods realise a mapping among nodes and point of the embedding space so that geometric relationships among embedded objects reflect the structure of the original graph. Since embeddings are points of an euclidean space, they may be used in other machine learning tasks (e.g. node classification) or in other graph analysis algorithms.

Currently, there exists many algorithms and many classification attempts that are categorised and described in some previous surveys [2,10,8,4,6,1].

The input of representation learning algorithms is a undirected and unweighted graph $G = (V, E)$ with its associated adjacency matrix A and a real-valued matrix X containing node attributes $X \in R^{mx|V|}$. The goal of each algorithm is to map each node into a vector z $\in R^d$ where $d < |V|$.

Shallow embedding methods encode each node ($v_i \in G$) into a single vector through the use of a simple encoding function defined as:

$$ENC(v_i) = Mv_i \tag{3}$$

where M is a matrix containing the embedding vectors and v_i is a vector used for selecting the column. The matrix M contains all the embeddings.

Each column of M encodes a node of the original graph and the number of rows d is lower than the number of nodes n. These embeddings were initially inspired by matrix factorization approaches. The differences among these methods are in the use of different loss function and similarity measures.

4 Workflow of the Experiment

In this work we focused on the comparation between a direct analysis of PCNs using network clustering and network embedding approach followed by clustering on the embeddings. We used PCN-Miner, implemented in Python 3.8 programming. It uses scypy and numpy libraries for managing matrices; management of

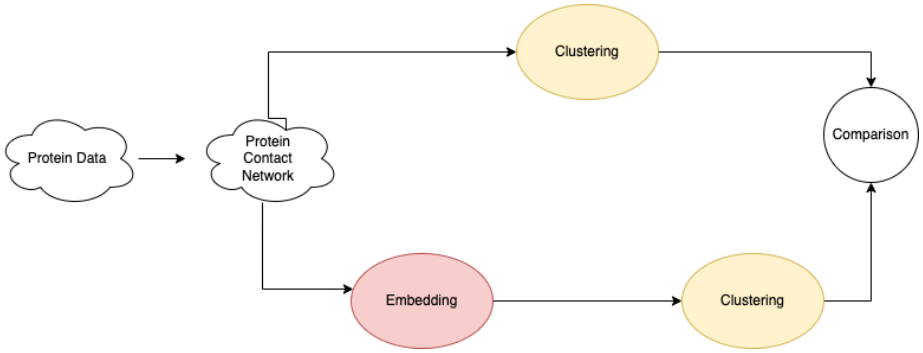

Fig. 1: Workflow

PDB files is provided by ProDy package; the network embedding is realised by wrapping the GEM library and clustering algorithms bu CdLib; visualisation of protein structures is made by wrapping the community edition of PyMol. Analysis started from protein data. The reference database for protein structures is the Protein Data Bank (PDB `https://www.rcsb.org/`), which also defines the PDB format, a standard for recording atom files. Using PDB files we obtain protein contact networks.

First step consist of import protein structure in PDB format from which it is possible to obtain the PCN (alternatively we can directly import a PCN previously determined). After we access analysis fuctionalities. On one hand we work with network clustering, on the other we work with network embeddings followed by clustering on the embeddings. Therefore we compared the results by comparison of centrality measures and participation coefficient, computed for each residue.

We focused on four structures of Spike protein, in its closed form: the wild type and three variants of concern (alpha, delta, omicron).

Figure 2 shows the Structure of the SARS-CoV-2 spike glycoprotein pdb code 6VXX. We first built the protein contact network using PCN-Miner. Then we embedded the resulting graph by using the HOPE algorithm. Each node was embedded into a vector having 64 dimension. Finally we applied the spectral clustering algorithm. We found some interesting community that could be annotated as allosteric regions after verification. Figure 3 shows the Structure of the SARS-CoV-2 spike glycoprotein (closed state)(pdb code 6VXX). This structure is the result of the clustering analysis, with soft clustering using a normalised lapacian. Figure evidences the found communities.

Figure 4 shows the structure of Closed state of pre-fusion SARS-CoV-2 Delta variant spike protein (pdb code 7SBK). The structure is the result of the embeddings+clustering analysis, with HOPE as embdeddings algorithm.

Figure 5 shows the Structure of the SARS-CoV-2 spike glycoprotein (closed state)(pdb code 6VXX). Closed state of pre-fusion SARS-CoV-2 Delta variant

 doi:10.20944/preprints202204.0027.v1

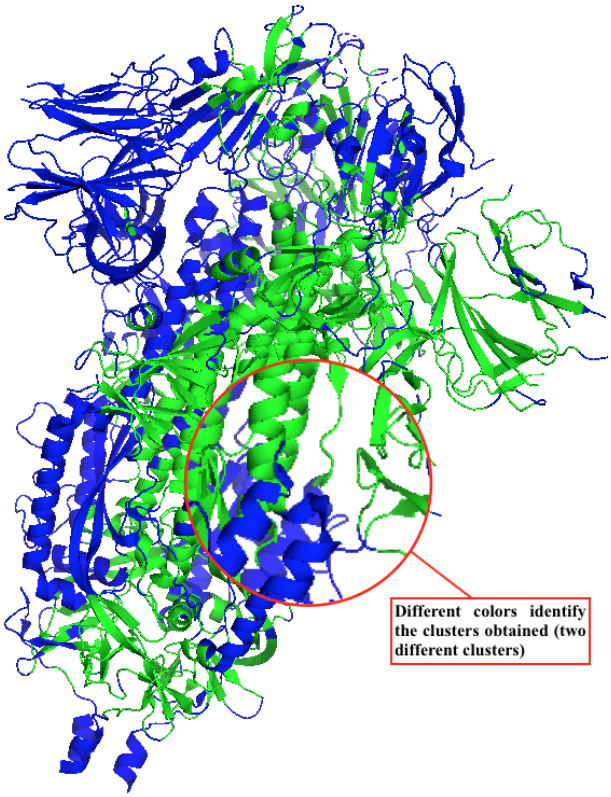

Fig. 2: Structure of the SARS-CoV-2 spike glycoprotein (closed state)(pdb code 6VXX). This structure is the result of the embeddings+clustering analysis, using the HOPE for node embedding

spike protein (pdb code 7SBK). This structure is the result of the clustering analysis, with soft clustering using a normalised lapacian.

 doi:10.20944/preprints202204.0027.v1

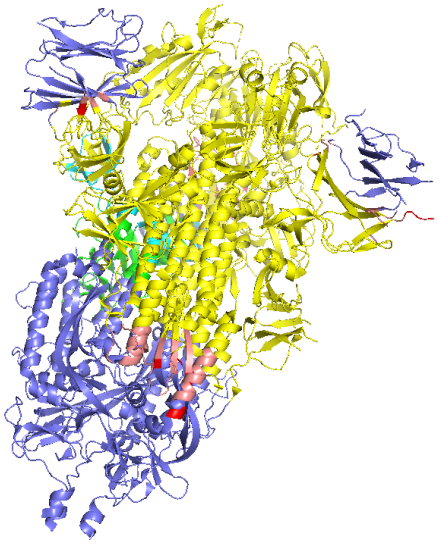

Fig. 3: Structure of the SARS-CoV-2 spike glycoprotein (closed state)(pdb code 6VXX). This structure is the result of the clustering analysis, with soft clustering using a normalised lapacian.

 doi:10.20944/preprints202204.0027.v1

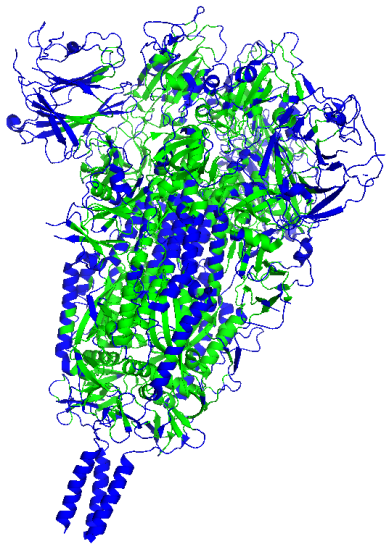

Fig. 4: Structure of Closed state of pre-fusion SARS-CoV-2 Delta variant spike protein (pdb code 7SBK). This structure is the result of the embeddings+clustering analysis, with HOPE as embdeddings algorithm.

5 Conclusion

Protein Contact Network (PCN) is an emerging paradigm for modelling protein structure. A common approach to interpreting such data is through network-based analyses. It has been shown that cluster- ing analysis may discover allostery in PCN. Nevertheless Network Em- bedding has shown good performances in discovering hidden communities and structures in network. In this work, we compare some approaches for graph embedding with respect to some classical clustering approaches for annotating protein structure

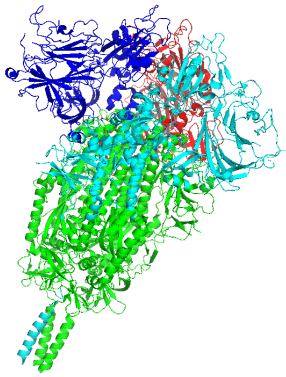

Fig. 5: Structure of Closed state of pre-fusion SARS-CoV-2 Delta variant spike protein (pdb code 7SBK). This structure is the result of the clustering analysis, with soft clustering using a normalised lapacian.

6 Acknowledgements

B.P. and U.L. are funded by PON-VQA Annotating Query Answering.

References

1. Cannataro, M., Guzzi, P.H., Sarica, A.: Data mining and life sciences applications on the grid. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **3**(3), 216–238 (2013)
2. Cui, P., Wang, X., Pei, J., Zhu, W.: A survey on network embedding. IEEE Transactions on Knowledge and Data Engineering **31**(5), 833–852 (2018)
3. Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., Giuliani, A.: Protein contact networks: an emerging paradigm in chemistry. Chemical reviews **113**(3), 1598–1613 (2013)
4. Goyal, P., Ferrara, E.: Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems **151**, 78–94 (2018)
5. Guzzi, P.H., Di Paola, L., Giuliani, A., Veltri, P.: Design and development of pcn-miner: A tool for the analysis of protein contact networks. arXiv preprint arXiv:2201.05434 (2022)
6. Hamilton, W.L., Ying, R., Leskovec, J.: Representation learning on graphs: Methods and applications. arXiv preprint arXiv:1709.05584 (2017)
7. Khan, T., Ghosh, I.: Modularity in protein structures: study on all-alpha proteins. Journal of Biomolecular Structure and Dynamics **33**(12), 2667–2681 (2015)
8. Nelson, W., Zitnik, M., Wang, B., Leskovec, J., Goldenberg, A., Sharan, R.: To embed or not: network embedding as a paradigm in computational biology. Frontiers in genetics **10** (2019)
9. Paola, L.D., Mei, G., Venere, A.D., Giuliani, A.: Disclosing allostery through protein contact networks. In: Allostery, pp. 7–20. Springer (2021)

10. Su, C., Tong, J., Zhu, Y., Cui, P., Wang, F.: Network embedding in biomedical data science. Briefings in bioinformatics **21**(1), 182–197 (2020)
11. Tasdighian, S., Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., Palumbo, P., Mei, G., Di Venere, A., Giuliani, A.: Modules identification in protein structures: the topological and geometrical solutions. Journal of chemical information and modeling **54**(1), 159–168 (2014)