Protein-Protein Interaction Networks: A computational perspective
Report for SC3.101: Introduction To Biology

Raunak Seksaria : 2023113019

3rd December, 2024

Abstract

This report looks at PPINs, especially at their relevance in the context of cellular pathways, and how proteins interact with each other to regulate and bring about functions. Biochemistry of cellular pathways can be studied in detail with the help of PPINs. The report starts with the basics of network biology, representations of proteins as nodes in a graph, and the advantages with that approach, and focuses on why biological networks are different from random networks. There were 2 major classical experimental methods that could help reconstruct PPINs, the Y2H method, and the Mass Spectrometry method. However, with the advent of computational models, we can also use them to do high-throughput proteomic analysis. This report explores 2 of these methods: SVMs which are used to classify native and non-native like interfaces that are predicted by computational models, by using a classification algorithm. Other prediction methods involve various deep learning methodologies, like Graph Neural Networks, which has recently gained traction. In the end, the report summarises a few applications of PPINs.

1 Introduction

Protein–protein interaction (PPI) refers to the covalent(or rarely, even non-covalent) bindings of two or more proteins as responses to various signals or disturbances or stresses, (collectively called dysregulation) whichever apply. Based on such PPINs, more systematic protein networks were established gradually, known as the protein-protein interaction network (PPIN). Part of network biology, PPINs have made considerable impact on the study of many diseases, like cancer, neurodegenerative diseases (Alzheimer's, Parkinson's, etc.), and even cardiovascular diseases, because these diseases are also associated with the dysregulation of cellular pathways. PPINs help to identify how proteins interact within these pathways and how alterations in such interactions can lead to disease. For example, the misfolding or aggregation of proteins in Alzheimer's can and has been studied through PPINs to understand how these processes interfere with normal cellular functions.

Basic terminology

1. **Genome** : The complete set of DNA (genetic material) in an organism.
2. **Proteome** : The proteome is an expression of an organism's genome.
3. **Interactome** : The interactome is defined as the entirety of interactions between biological entities in a cell.
4. **Interlogs** : Interologs are protein-protein interactions that are conserved across different species, especially when the proteins in consideration have the same evolutionary origin.

2 Network Biology basics:

In network biology, we essentially think of representing intracellular biological networks via a graph. Typically this considers molecular components within a cell as nodes and their direct or indirect interactions as links. These macromolecules can be amino acid residues (in the context of protein contact networks, that can be used to study structural and intramolecular protein interactions), metabolites (in the context of metabolic networks, that represent a particular biochemical process in its entirety), drugs(in the context of drug interaction networks, where drugs are linked to their targets), and many more. Here we talk about one such kind of network, the protein-protein interaction network(PPIN).

2.1 Graph Theory Basics

A graph G is a mathematical discrete structure that is typically represented as a 2 tuple (V,E), where V represents the set of vertices(also called nodes), and E represents the set of edges(an edge connects two vertices). For each vertex, a term called degree is defined, which is the sum of the number of edges coming into(separately called indegree) and the number of edges going to(separately called outdegree). In essence, graph edges can be directed, or undirected(based on whether the edges have a set direction associated with them), and the terms indegree and outdegree (as written above) are used only in the directed context. We also note that a path between any two nodes is a series of edges, that if traversed, would connect the two nodes. Naturally, there could be pairs of nodes, that would not have any path between them. Moreover, our graph can be modified to include weights for each edges, which represent in our case a quantification of the strength, reliability, or significance of the interaction between the two proteins. In such cases, shortest paths, or the sequence of edges between any 2 nodes that have the lowest sum of weights becomes relevant.

A subgraph S(V_s, E_s is a subset of the original graph G(V,E) such that $V_s \subseteq V$ and $E_s \subseteq E$, all the while maintaining the original graph connections, in the particular subset of nodes and edges considered. Not all subgraphs occur with equal frequency. Some, known as motifs, are over represented as compared to the same subgraph's frequency in a random network. In particular a feed forward triangular subgraph({A,B,C},{$e_{A\to B}, e_{A\to C}, e_{B\to C}$}) is a motif regularly seen in biological networks. We also note that we can form clusters among the nodes that are present close enough to each other, representing high density areas of interaction.

2.2 Properties specific to biological networks

Most biological networks are scale-free. The scale-free property indicates that there are few very highly-connected nodes in the network. What makes this property especially useful in the context of biological networks, is the fact that such a network can tolerate a random protein removal, but is extremely sensitive to the removal of hub (hubs are the proteins that are extremely well connected[highest degree nodes]). The degree distribution approximates a power law, $P(k) \sim k^{-\gamma}$, where γ is the degree exponent, and $\sim$ represents 'proportional to', this being the formal definition of a scale-free network. The smaller the value of γ, the more important the role of the hubs in the network. The difference is highlighted in Figure 1.

Such graphs also typically have the small-world property. The small-world property indicates that any two nodes in a network can be connected with a small number of links, while the average path length between nodes becomes much shorter than a random network due to the existence of hub proteins(many are connected to it, and hence shortest paths usually include the hub node). This short path length indicates that local perturbations in concentrations or presence or absence of particular proteins in the network, could very quickly reach, and impact the entire network. The primary purpose of this description was to make sure that we understand that biological networks are not precisely like random networks, like the ones dealt with in mathematics. [1]

3 Classical Techniques used in PIN Reconstruction

3.1 Yeast Two-Hybrid(Y2H)

This technique identifies protein-protein interactions within living cells, relying on splitting a transcription factor into two parts: a DNA-binding domain (BD) and an activation domain (AD). These domains are attached to two proteins of interest, called "bait" and "prey." In Yeast Two-Hybrid (Y2H), the bait is the protein of interest fused to the BD, while the prey is the potential interacting protein fused to the AD. When the bait and prey interact, the functional transcription factor reassembles, activating a reporter gene.

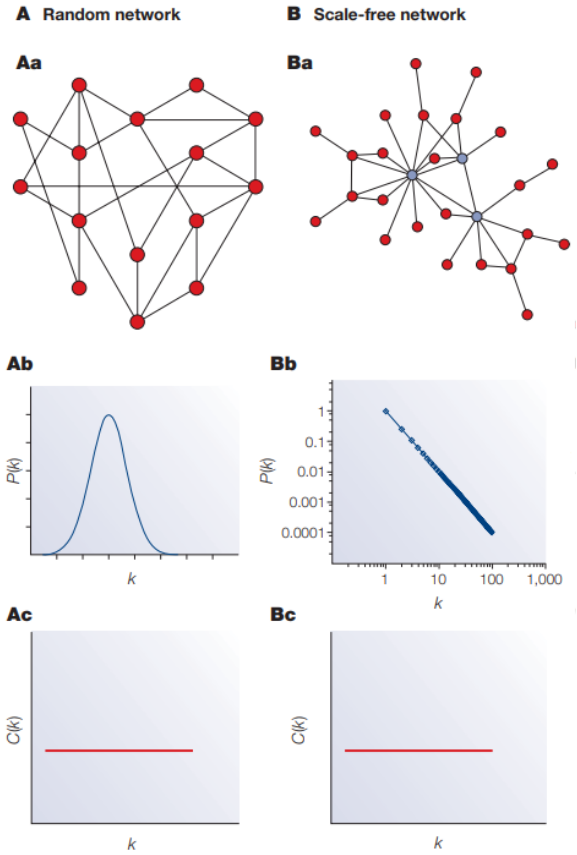

Figure 1: Shows the difference in biological(scale-free) and random networks. C(k) is the clustering coefficient, and P(k) is the degree distribution:probability that a specified node has exactly k links

It is a high-throughput method applied in the discovery of protein-protein interactions *in vivo*.High-throughput means that the method can test many protein combinations simultaneously, making it efficient for large-scale studies. It being *in vivo* means that the experiments occur within the living cells itself, allowing the interactions to be studied in a context that closely mimics natural biological conditions. Y2H is simple and a cost-effective approach is widely used for its ability to detect weak or transient interactions. However, Y2H has a high false-positive rate, so additional validation is often necessary. Many PINs were reconstructed based on Y2H, including PINs of E. coli and H. Sapiens, among many others. It was the most prevalent technique up until 2006.

3.2 Mass Spectrometry and Affinity Purification

Mass spectrometry-based methods are powerful tools for identifying protein-protein interactions, often combined with techniques like Affinity Purification (AP) and Tandem Affinity Purification (TAP). These methods capture proteins that interact with a ”bait” protein, which is immobilized using molecular tags or antibodies. Captured proteins, known as ”prey,” are then digested into smaller fragments (peptides) and analyzed by mass spectrometry, which identifies their molecular weight and structure. TAP improves specificity by using two rounds of purification, while co-Affinity Purification (coAP) uses antibodies instead of tags, minimizing disruptions to protein structure. These approaches are largely performed in vivo as well. One of the key advantages of mass spectrometry-based methods is their ability to detect post-translational modifications (such as phosphorylation) that occur in living systems, providing insights into proteins in their natural states. However, they are less effective at detecting weak or transient interactions and may introduce false negatives if the tags interfere with protein folding. Overall, these methods combine higher accuracy with a focus on biologically relevant conditions, making them relevant in the context of mapping protein networks and understanding cellular processes. [3]

4 Use of Computational Tools

4.1 Graph Neural Networks(GNNs):

In the paper referred, a deep learing method was used that combined graph neural network (GNN) and a language model (LM) to predict the interaction between proteins. The LM was used to generate per-residue embedding from a PDB sequence, which was used as the node's feature of the protein graph. The GNN-based model then extracted features from the protein's graphical representation (combining structural and sequence information). The outputs of the GNN model were fed to the PPI classifier. Two models were used, both of which are described below, and represented in Figure 2. [4]
The unique capacity of GNNs to exploit the underlying structure of graph data is achieved through several key mechanisms. Firstly, GNNs can use neighborhood aggregation or message-passing frameworks, wherein each node in the graph gathers information from its local neighbors to update its state. This allows GNNs to incorporate local context into node representations, important because of the fact that PPINs have motifs as their primary building block, thereby capturing the immediate interaction dynamics in PPINs. Secondly, through multiple rounds of these aggregations, GNNs can learn increasingly abstract representations of nodes, thereby modeling higher-order interaction effects and uncovering complex interaction patterns. [6]

4.1.1 Graph convolution network:

Let $P = (V, E)$ be a graph representation of a protein, such that $|V| = L$, L being the length of the PDB sequence. Each layer of the GCN takes the adjacency matrix $A \in R^{L,F}$(where F is the dimension of node features) and node embeddings from the previous layer $H^{(l)} \in R^{L,F_l}$ as input and outputs the node-level embeddings for the next layer $H^{(l+1)} \in R^{L,F_{l+1}}$. Kipf and welling defined a particular equation for this:

$$H^{(l+1)} = \text{ReLU}\left(\hat{D}^{-0.5}\hat{A}\hat{D}^{-0.5}H^{(l)}W^{(l+1)}\right)$$

Here, $\hat{A}$ is the adjacency matrix added with identity matrix $I_L \in R^{L,L}$ ($\hat{A} = A + I_L$). The addition of identity matrix to adjacency matrix enforces self-loops in the graph. $\hat{D}$ is the diagonal node degree matrix calculated as: $\hat{D}_{ii} = \sum_{j=1}^{L} \hat{A}_{ij}$. It is used to normalize the adjacency matrix . Also ReLU is a standard activation function used in neural networks. Another standard activation function, used in Graph attention networks below, is the leaky ReLU function.

4.1.2 Graph attention network

This, also called GAT is an attention-based architecture that operates on graph-structured data. The idea is to use a self-attention method to compute hidden representations of each node in the graph by looking at its neighbors. In GAT, for input feature matrix $X \in R^{L,F}$, we will get the learned feature matrix $H \in R^{L,F'}$ as an output. Here, L is the number of nodes in the graph, and F and F' are the input and output dimensions, respectively, just as written above. The expression to calculate the learned feature of node i is:

$$H_i = \sigma\left(\sum_{j \in N_i} \alpha_{ij} W X_j\right)$$

Here, σ is the Leaky ReLU activation function, and neighbouring nodes are represented as N_i. W is the weight matrix, and normalization constant(just written in general) is written as $\alpha_{i,j}$, which can be calculated.

4.2 Use of Support Vector Machines(SVMs):

4.2.1 Support Vector Machines:

Support Vector Machines (SVMs) are supervised ML algorithms used in various classification and regression models. The core idea is to find the optimal hyperplane that separates data points into different classes in a high-dimensional space, based on the possible features. SVMs use kernel functions (e.g., linear, polynomial, radial basis function) to transform input data into a form where separation is easier. The algorithm focuses on maximizing the margin between classes, ensuring robust and generalized predictions.

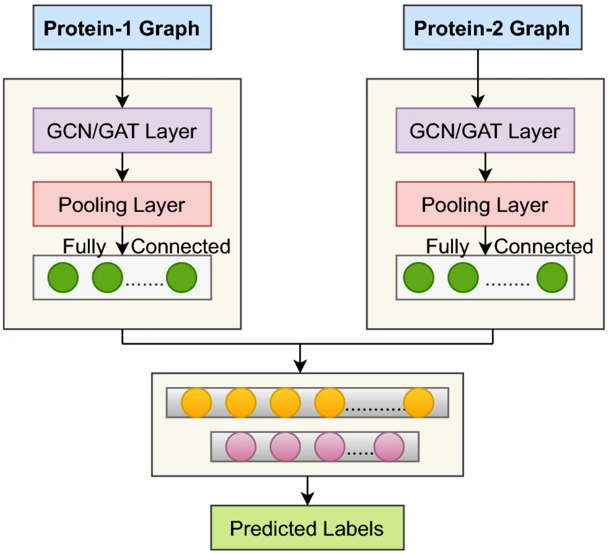

Figure 2: An example GNN model

4.2.2 Use of SVMs in PPIN:

The model used in the paper referred here is a Support Vector Machine (SVM)-based classification model designed to distinguish between native-like and non-native protein-protein interaction (PPI) interfaces. It utilizes a standard radial basis function (RBF) as the kernel function. The input to the model consists of structural and chemical properties of protein interfaces, such as the accessible surface area (ASA), buried surface area (BSA), hydrogen bonds, salt bridges, and binding energy. The dataset comprises native-like protein complexes derived from experimental data and non-native complexes generated via computational docking. Docking means generating potential interaction arrangements by exploring translational, rotational, and conformational spaces and scoring them using algorithms to identify the most favorable binding mode for interacting molecules. The non-native complexes are further categorized based on thresholds for metrics like fraction of conserved native contacts (FNAT) and interface root mean square deviation (iRMSD). A high FNAT means a large overlap between the residues forming the interface in the docked model and the reference native structure, whereas a low iRMSD means that the spatial arrangement of residues in the docked interface is close to that of the native complex. These thresholds enable classification into highly, moderately, and weakly distinguishable non-native complexes. [2]

A native-like interface in the context of protein-protein interactions (PPINs) refers to an interaction surface between two proteins that closely resembles the interface found in the biologically relevant or experimentally determined protein complex structure. This is especially useful, because it resembles which of our findings resembles *in vivo* methodologies. Native-like interfaces also typically exhibit favorable binding free energy, indicating a stable and functional interaction. In reality, non-native interfaces are biologically irrelevant predictions.

4.3 Other deep learning methods

Recurrent Neural Networks (RNNs) are neural networks designed to recognize patterns in sequences of data, in this case, protein sequences. They“remember” previous inputs in the sequence using hidden states, making them useful for PPIN analysis. CNNs, transformer models with attention mechanisms, autoencoder models, are also a few of the other deep learning methodologies used in PPIN analysis and reconstruction

5 Analysis and Applications of PPIN

The strategies that aim to interpret PPINs generally try to find common attributes within members of the network. Useful attributes include similarities in expression pattern, function or subcellular localization. Nodes may also be grouped on the basis of network topology: groups of highly interconnected nodes may form clusters. [5]

5.1 Functional Annotation of Proteins

Although we know a lot about the functions of proteins, with advances in technologies, every day we find new proteins that we dont know a lot about. Functional annotation refers to identifying the functional aspects of every protein. PPIns help in such kinds of predictions for unknown proteins by analyzing their interactions with proteins who are aldready functionally annotated(known functions). Functions can be inferred based on neighboring proteins' roles or through clustering-based analysis, because proteins in the same cluster often share similar functions. This method is a rapid, high-throughput alternative to traditional experimental annotation and is especially valuable for preliminary annotation in novel proteins.

5.2 Substem Investigation and Pathway Mappings:

Subsystem analysis involves extracting specific sub-networks of the PPIN(similar concept like subgraph in section 2.1) to study certain extremely specific biological pathways in detail. These sub-networks focus on groups of proteins associated with specific functions(which can mean looking at specific clusters in the network), and map them as DNA repair, immune response, or signaling pathways. Subsystem analysis can also highlight connections between diseases, as seen in cancer-related proteins forming tightly connected sub-networks, that have been linked to tumorigenesis.

5.3 Evolutionary Analysis

PPINs are used to study evolutionary relationships by identifying interlogs across species. They reveal how new cellular functions emerge through interactions between ancient and newly-evolved proteins. Co-evolution patterns between interacting proteins can also help predict new interactions and trace evolutionary paths of biological processes. The idea is that most PPINs have a common pathway for most functions across many species. For example, in cancer research, there is a common general network of proteins that suppress or activate cancer growing abilities of the cell. Proteins having similar evolutionary origins can be researched using this common network, and minor modifications made for their specific pathways, which is particularly useful in large-scale proteomics.

6 Conclusion

Although many organisms have been comprehensively studied for their protein-protein interactions, none of the PINs are capable of capturing all of the interactions in the cell. In fact, for most studies, the proteins detected with interactions usually cover no more than 30% of the whole proteome, which indicates the broader development potentiality of PIN research. One of the key future emerging areas involves the prediction of specific aspects of PPINs beyond merely identifying whether an interaction occurs. This includes predicting the binding sites of PPINs, understanding residue-residue interactions across protein interfaces, and determining protein–protein association rates. One of the biggest current models for PPIN analysis is AlphaFold2.

With the accumulation of PPI information obtained from different methods, a series of databases have been established to collect and manage the PPINs from various organisms. However, the amount of data available hardly compares to the kind of biological networks for each organism that we can have. An insight into PPINs gives us fundamental biochemical knowledge, and is crucial for drug design and systems biology approaches for diseases. Hence, further research into the topic, and usage of the datasets that are already available in functional annotation, evolutionary and disease analysis, and in subsytem investigation, is the way forward for the field.

References

[1] Albert-László Barabási and Zoltán N. Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. `doi:10.1038/nrg1272`.

[2] Subhrangshu Das and Saikat Chakrabarti. Classification and prediction of protein–protein interaction interface using machine learning algorithm. *Scientific Reports*, 11(1):1761, 2021. Available at: `http://www.hpppi.iicb.res.in/pcpip/`. `doi:10.1038/s41598-020-80900-2`.

[3] T. Hao, W. Peng, Q. Wang, B. Wang, and J. Sun. Reconstruction and application of protein–protein interaction network. *International Journal of Molecular Sciences*, 17(6):907, 2016. doi:10.3390/ijms17060907.

[4] K. Jha, S. Saha, and H. Singh. Prediction of protein–protein interaction using graph neural networks. *Scientific Reports*, 12(1):8360, 2022. doi:10.1038/s41598-022-12201-9.

[5] Gavin C. K. W. Koh, Pablo Porras, Bruno Aranda, Henning Hermjakob, and Sandra E. Orchard. Analyzing protein–protein interaction networks. *Journal of Proteome Research*, 11(4):2014–2031, 2012. PMID: 22385417. doi:10.1021/pr201211w.

[6] M. Lee. Recent advances in deep learning for protein-protein interaction analysis: A comprehensive review. *Molecules*, 28(13):5169, 2023. doi:10.3390/molecules28135169.

[7] Qianmu Yuan, Jianwen Chen, Huiying Zhao, Yaoqi Zhou, and Yuedong Yang. Structure-aware protein–protein interaction site prediction using deep graph convolutional network. *Bioinformatics*, 38(1):125–132, 09 2021. doi:10.1093/bioinformatics/btab643.