Graph Representations and Algorithms in Biomedicine

Brianna Chrisman[1], Maya Varma[1], Sepideh Maleki[2], Maria Brbic[3], Cliff Joslyn[4], Marinka Zitnik[5]
[1]Stanford University, [2]UT Austin, [3]EPFL, [4]Pacific Northwest National Labs,[5]Harvard University

1. Introduction

Connectivity is a fundamental property of biological systems: on the cellular level, proteins interact with each other to form protein-protein interaction networks (PPIs); on the organism level, neurons are arranged in a network; and on a community-level, species can have complex relationships with one another that drive the development and balance of an ecosystem. Graphs, representations of systems consisting of entities as vertices and their connections as edges, are a useful structure to characterize many such systems. Such models can be used to understand biological systems that naturally have a network structure, including PPIs, biological neurons, and ecosystems. In today's information age, graph representations and algorithms (often in combination with machine learning techniques) are used to organize massive amounts of related data, much of which may be heterogeneous or unstructured, and identify patterns that represent novel biological insights. PSB's 2023 session ``Graph Representations and Algorithms in Biomedicine," encompasses modern developments in graph theory and its applications to various fields of biomedicine. This session includes a wide range of research - knowledge graphs built from text-mined health data, heterogeneous networks using multi-omic databases, and graphs refined to represent uncertainty or improve memory usage.

Recent developments around graphs in biomedicine have primarily revolved around methods of constructing, comparing, and making predictions from graphs using massive datasets that have become commonplace in biomedical computation. Even more challenging, or perhaps more opportune, is that many problems in biomedicine involve multiple different data types. A specific challenge is how to integrate heterogeneous, sometimes unstructured data, to make network-based insights. The proceedings for this session tackle several different challenges: understanding and predicting protein networks (Eyuboglu et al., Ayati et al.), improving feature representations of various types of graphs (Chen et al., Soman et al., Luo et al.), making use of family structure via graph approaches (Shemirani et al., Mossel et al.), creatively applying traditional algorithms to novel tasks (Magnano et al.), and representing uncertainty in network structures (Liu et al., Krishnan et al.).

2. Understanding and Predicting Molecular Networks

Predicting the structure, function, and associated phenotypes of molecular networks has emerged as a grand challenge that is very amenable to graph-based approaches. One past related strategy for protein-protein interaction network prediction has been to quantify protein similarly in terms of protein sequence similarity, or their distance to one another in the network. However, Eyuboglu et al., Ayati et al., illustrate that there are other factors we can consider when trying to predict protein or molecule similarities, phenotypes, and networks. In their paper, "Mutual Interactors as a Principle for the Discovery of Phenotypes in Molecular Networks", Eyuboglu et al. suggest that molecular similarity is not dictated by molecule-molecule distances in graph space, but is better described using representations of a molecule's mutual interactors. They show that this principle - that molecules with similar sets of mutual interactors have similar phenotypes - holds for protein-protein, signaling, and genetic networks. To further showcase the application of this theory in practice, they build a machine learning model using a simple mutual interactor feature space, and illustrate that they can predict drug targets, disease proteins, and molecular functions better than complex algorithms and feature spaces.

Interestingly, Ayati et al. take a comparatively opposite approach. They argue that while many past strategies to predict kinase-substrate associations have used sequences alone, there is a wealth of publically available information on protein structure and function that could vastly improve kinase-substrate predictions. The authors use sequence similarity, shared molecular pathways, and co-evolution, co-occurrence, and co-phosphorylation patterns to construct a phosphosite-phosphosite association network, and protein-protein interactions, mutual biological pathways, and kinase family membership to construct kinase-kinase networks. Using these networks to represent kinase and substrates' node embeddings, they train a machine learning model that outperforms the state-of-the-art methods for predicting kinase-substrate interactions. Ayati et al.'s complex node embeddings using heterogenous information sources, and Eyuboglu et al.'s simple and interpretable representations of molecular similarities illustrate two different and creative approaches for improving the feature space that we use to understand and make predictions on molecular networks.

3. Understanding and Predicting Molecular Networks

Key contributions in both Ayati et al and Eyuboglu et al were improved representations of the feature space of molecular networks. Improving network feature representations - reducing memory or runtime requirements, boosting interpretability, or increasing accuracy in downstream machine learning pipelines - is a general goal of research in biomedical networks. "Contrastive learning of protein representations with graph neural networks for structural and functional annotations" by Luo et al., “A Graph Coarsening Algorithm for Compressing Representations of Single-Cell Data with Clinical or Experimental Attributes" by Chen et al., and "Time-aware Embeddings of Clinical Data using a Knowledge Graph" by Soman et al. all tackle this challenge in various ways.

In "Contrastive learning of protein representations with graph neural networks for structural and functional annotations", like Ayati et al and Eyuboglu et al, Luo et al. focus their efforts on the protein space. Rather than trying to use functional and structural information to predict protein-protein interactions, they use the ladder to predict functional and structural annotations. Their algorithm, "PenLight" uses a graph neural network (GNN) that integrates three dimensional protein structure, and sequence representation using a language model. They use contrastive earning to train the GNN to learn protein representations that reflect similarities encompassing not only similarities in the linear sequence space, but semantic similarities and similarities in the function or sequence space. They benchmark their algorithm on predicting EC (Enzyme Commission) numbers and CATH (class, architecture, topology, homologous superfamily) classifications, functional and structural annotations respectively available on the Protein Databank, demonstrating its superior performance.

In "A Graph Coarsening Algorithm for Compressing Representations of Single-Cell Data with Clinical or Experimental Attributes", Chen et al. introduce a novel approaching for compressing graphs of single-cell data. In single-cell experiments, measurements from tens or hundreds of thousands of cells are often visualized and analyzed by looking at a dimensionality-reduced representation of the cells. This dimensionality reduced representation of the cells can also be described in a graph, where cells or groups of cells with similar features in the latent space are connected to each other on the graph. Chen et al. develop a method for performing graph coarsening on this network, which can save memory, remove noise, and help distinguish biologically relevant patterns in downstream pipelines. Importantly, their algorithm “cytocoarsening”} not only uses not only cell-cell similarity in the single-cell measurements (in their case they were using mass cytometry data), but also clinical, experimental, and phenotypical attributes of the cells. Using single cell mass cytometry datasets from cohorts from studies of preeclampsia, COVID-19, and cytomegalovirus, the authors demonstrate that their algorithm has comparable runtime to state-of-the-art graph coarsening packages, and improved performance when it comes building coarsened graphs that depict biologically relevant patterns.

Finally, in "Time-aware Embeddings of Clinical Data using a Knowledge Graph", Soman et al. construct biomedical knowledge graphs from electronic health records to create machine readable representations of patient health data. They map a patient EHR data onto nodes of a popular biomedical knowledge network and use a random walk to create node embeddings with features corresponding to nodes in the knowledge network graph. To capture to temporal dynamics of the EHR data, they build embedding vectors unique to each yearly interval time frame. Such embeddings yield a highly interpretable two-dimensional array, with rows representing time and columns representing SPOKE nodes. Using these embeddings as feature representations for patients from a group of Parkinson's and non-Parkinson's phenotypes, they build a machine learning model that can predict Parkinson's using data from one year or earlier before a patient's diagnosis. Feature representations without the temporal representation were not as predictive, illustrating that the dynamic nature of electronic health records is an important aspect to capture when creating feature representations of EHR data.

4. Making Use of Family Structure

While molecular networks are an obvious candidate for graph representations and algorithms, another candidate is genetic data from related individuals. A classic family tree is a graph, and graphs can also depict more complicated genetic relationships from individuals. In "Selecting clustering algorithms for Identity-by-descent mapping" by Shemirani et al. and "Efficient Reconstruction of Stochastic Pedigrees: Some Steps from Theory to Practice" by Mossel et al., both authors use graphs to understand and quantify the genetic relatedness of individuals.

In "Selecting clustering algorithms for Identity-by-descent mapping" Shemirani et al. develop a metric for benchmarking identity-by-descent clustering algorithms. They introduce a novel approach for finding groups of individuals that share short segments of their genome inherited from a recent common ancestor (a concept known as "identical-by-descent"). They designed a clustering benchmark and used it to compare the performance of several popular IBD clustering algorithms. They found that Infomap and Markov clustering community detection methods had the highest statistical power in finding communities with shared IBD. Notably, they show that traditional clustering metrics, such as modularity and purity, do not necessarily provide the highest statistical power to IBD clustering applications, necessitating the development of improved IBD clustering benchmarking strategies.

In "Efficient Reconstruction of Stochastic Pedigrees: Some Steps from Theory to Practice", Mossel et al. build on their previous work where they reconstructed a pedigree from genetic data under a number of simplifying assumptions. In this newer work, the authors walk us through the process by which they made simplifications to improve the runtime of their algorithm, observe scenarios in which the faster algorithm has decreased performance, identified the theoretical issues and limiting cases with their new approach, and correct accordingly. Specifically, they found that the faster version of their algorithm performs with pedigrees that are beyond 2 generations. They claim that this is due to inbreeding nearly always present in large pedigrees, and show that the algorithm improves when inbreeding is limited in their simulation. Finally, they introduce a belief propagation heuristic that helps account for possible inbreeding, allowing for both fast and accurate pedigree reconstruction.

5. Applying Traditional Graph Algorithms to Novel Tasks

Molecular networks and pedigrees are natural structures by which graph strategies can be applies, but Magnano et al. show that traditional graph-based approaches can show promise for novel tasks. In "Graph algorithms for predicting subcellular localization at the pathway level", Magnano et al predict subcellular protein localization using an edge labeling task. Using biological pathway networks, the authors develop graph algorithms in order to predict the location within a cell that an interaction is taking place. They pose this challenge as an edge-labeling task and compare the performance of a variety of several models including GNNs, probabilistic models, and discriminative classifiers. Notably they found that directly using data from protein localization databases was not sufficient to accurately predict pathway level localization and topology or some other form of structural information is needed to predict localization in context. Finally, they use their findings to predict interaction localizations in a human cytomegalovirus infection.

6. Representing Uncertainty in Networks

A major weak point that often goes unaddressed in biomedical graph-related networks is that networks derived from publicly available data have noise and potential inaccuracies in their structures and topologies. Often this goes unaddressed, but accounting for such inaccuracies or better understanding their effects may allow us to build more graph-based feature representation and models of biological phenomenon. In "Improving target-disease association prediction through a graph neural network with credibility information", by Liu et al. and "Integrated Graph Propagation and Optimization with Biological Applications" by Krishnan et al., the authors tackle the challenge of representing uncertainty in such biological networks.

"Improving target-disease association prediction through a graph neural network with credibility information" Liu et al., hope to improve target-disease association (TDA) predictions using biological networks and text mined data from the literature. They develop creatTDA - a deep learning based framework that learned latent feature representations of targets and diseases. Uniquely, they propose a new way to encode credibility information obtained from literature in their mode. They do this by learning credibility encodings for different known target-disease associations, using their co-occurrences in the literature as a label. CreaTDA was able to predict known TDAs with higher sensitivity and specificity, as well as novel TDAs including an association between bronchiolitis and the epidermal growth factor receptor and viral diseases and vascular endothelial growth factor.

In "Integrated Graph Propagation and Optimization with Biological Applications," Krishnan et al. seek to understand how uncertainty effects graphs representing biological network dynamics. In mathematical models of biological systems, rate constants are often unknown and network propagation has emerged as a suitable method for understanding how changes in nodes effect one another, without the need for parameter estimation. Krishnan et al extend some of the ideas in network propagation theory to develop a system of identifying which specific perturbation patterns may drive networks into desired states. Their method Integrated Graph Propagation and Optimization (IGPON) embeds propagation into an objective function and uses optimization strategies to minimize the difference between a target and observed state. They illustrate the value of their algorithm on predicting gene expression patterns using various sets of knockout data.

7. Conclusion

This session of papers addresses a wide variety of biological challenges: predicting molecular interactions, deriving insights from unstructured EHR data, quantifying genetic relationships between related individuals, and understanding the relationships between drug, disease, and phenotype. Excitingly, these works tackle these challenges using a diverse collection of graph-based approaches. We hope the common language of graphs will make apparent the intersections and differences in the problems addressed and the strategies taken, and readers and authors alike will be able to take additional inspiration from the ideas posed in this session.

References

Ayati, M., Yilmaz, S., Lopes, F., Chance, M., Koyuturk, M. "Prediction of Kinase-Substrate Associations Using The Functional Landscape of Kinases and Phosphorylation Sites". Proceedings of the Pacific Symposium for Biocomputing 2022.

Chen, C., Crawford, E., Stanley, N. 1 "A Graph Coarsening Algorithm for Compressing Representations of Single-Cell Data with Clinical or Experimental Attributes". Proceedings of the Pacific Symposium for Biocomputing 2022.

Eyuboglu, S., Zitnik, M., Leskovec, J. "Mutual interactors as a principle for phenotype discovery in molecular interaction networks". Proceedings of the Pacific Symposium for Biocomputing 2022.

Krishnan, K., Shi, T. "Integrated Graph Propagation and Optimization with Biological Applications". Proceedings of the Pacific Symposium for Biocomputing 2022.

Liu, C., Yu, C., Lei, Y., Lyu, K., Tian, T., Li, Q., Zhao, D., Zhou, F., Zeng, J. "Improving target-disease association prediction through a graph neural network with credibility information". Proceedings of the Pacific Symposium for Biocomputing 2022.

Luo, J., Luo, Y. "Contrastive learning of protein representations with graph neural networks for structural and functional annotations". Proceedings of the Pacific Symposium for Biocomputing 2022.

Magnano, C.S., Gitter, A. "Graph algorithms for predicting subcellular localization at the pathway level". Proceedings of the Pacific Symposium for Biocomputing 2022.

Mossel, E., Vulakh, D. "Efficient Reconstruction of Stochastic Pedigrees: Some Steps From Theory to Practice". Proceedings of the Pacific Symposium for Biocomputing 2022.

Shemirani S. , Belbin G.M., Burghardt, K., Lerman, K., Avery, C.L., Kenny, E.E., Gignoux, C.R., Ambite, J. "Selecting Clustering Algorithms for Identity-By-Descent Mapping". Proceedings of the Pacific Symposium for Biocomputing 2022.

Soman, K., Nelson, C.A., Cerona, G., Baranzini, S.E. "Time-aware Embeddings of Clinal Data Using a Knowledge Graph". Proceedings of the Pacific Symposium for Biocomputing 2022.