Graph Matching, Pattern Learning, and Protein Modeling

Wei Qian, Pengyu Hong | Computer Science Department, Brandeis University, Waltham, MA, USA

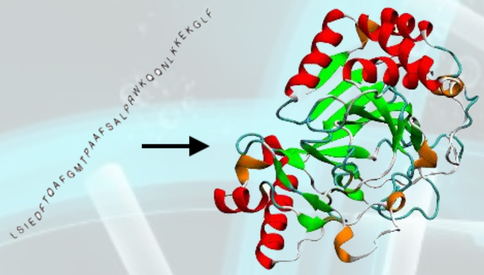
I. Introduction

One of the most amazing capabilities of human beings is to extract common spatial patterns from observations and use these patterns to make inference. For example, even you do not know Woody and if I tell you that below are two pictures of him, you will still be able to recognize him from the pictures since he is the only person (i.e. the common spatial pattern in our case) showing up in both pictures:

To represent these spatial patterns, we utilized a graph representation in computer science called ARGs (attributed relational graphs) consisting of labelled nodes connected by labelled edges:

and we build algorithms to compare different ARGs (this is called graph matching/graph alignment) and extract the common pattern (the **orange** part) while ignore the background noise (the **purple**, **black**, and **blue** part) from the ARGs (this is called pattern learning).

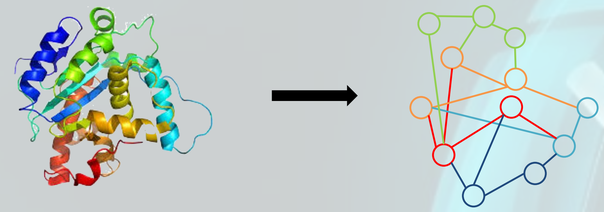
One important application of such algorithms would be modeling proteins in their 3D structures. Proteins are macromolecules responsible for nearly every task of cellular life. They are first generated by DNA (the poster background!) as a linear sequence of amino acids and then folded as a 3D structure in order to function:



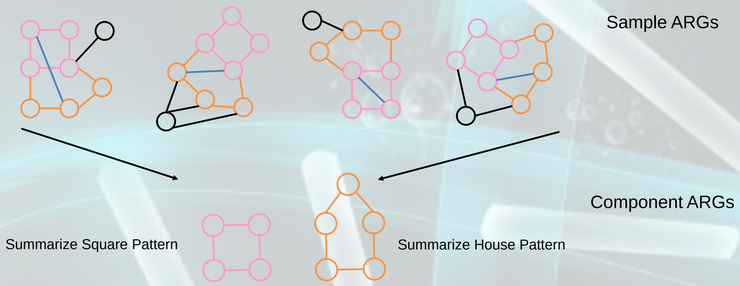
Conventional computational methods model proteins in their unfolded sequence form (shown on the left), and ignore the 3D nature of proteins (shown on the right). By modeling proteins in their 3D structures and taking advantages of the available 3D structure data, our algorithms have the potential to discover new structure patterns in proteins that are impossible to discover in their linear structure.

II. Method

To discover pattern in protein, we first turn their 3D structures into ARGs to represent their spatial relationship where nodes represent individual amino acid in the protein and edges represent a closed distance between the two amino acids:



Then to learn the common patterns among these ARGs, we used some small number of special ARGs as our component ARGs to represent the entire set of sample ARGs:



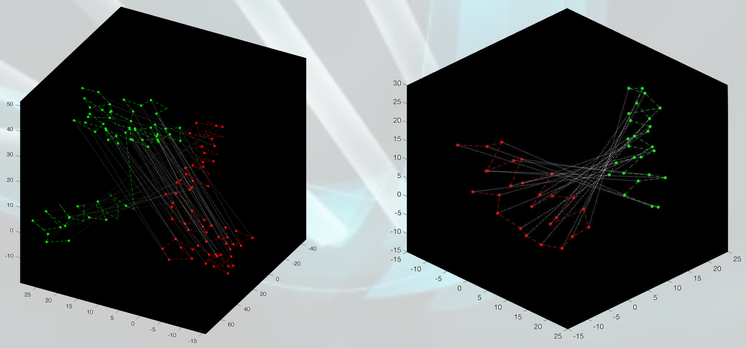
Theses component ARGs are initially picked from the ARGs set randomly, and updated iteratively in order to maximize the following energy function while matching to every ARG in the set:

$$E(G,G',M) = -\frac{1}{2}\sum_{a\in G}\sum_{b\in G}\sum_{i\in G'}\sum_{j\in G'} M_{ai}M_{bj}C_{abij} - \alpha\sum_{a\in G}\sum_{i\in G'} M_{ai}C_{ai} - \beta\sum_{a\in G}\sum_{b\in G} M_{a-1b-1}M_{ab}M_{a+1b+1} \quad (1)$$

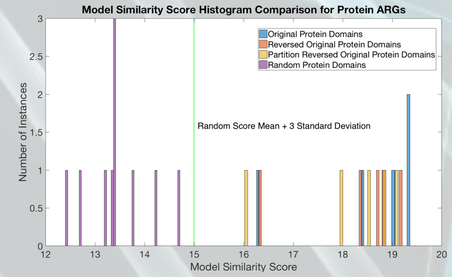
where M is the matching matrix from one graph to the other, and C is the edge and node similarity.

III. Result

Our matching algorithm based on eq.1 can generate correct matches between 3D structures of two proteins (red and green) shown below where very similar structures have brighter matching lines:



After feeding in proteins with CH1 domain, the trained components ARGs successfully capture the CH1 domain pattern, and easily distinguished the CH1 domain pattern from random protein pattern even when the pattern is reversed, or partitioned into a different order:

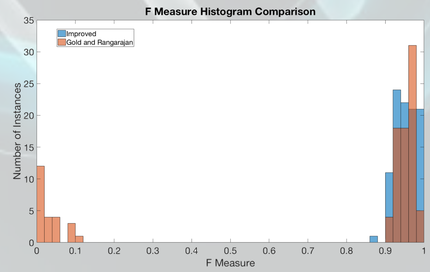


IV. Discussion

To compare two proteins, we used a graph matching algorithm developed by Gold and Rangarajan in 1996. However, their algorithm does not work well enough for large graphs like protein structures. Therefore, we made the following improvements result in a significant performance boost:

1) introduced random noises to avoid suboptimal solutions

2) introduced a null network to match background noise



Even though we match proteins by their 3D structure, when comparing two proteins, we use a local context match term shown in red for eq.1 in order to encourage the algorithm to match proteins sequentially and give us a clearer and more feasible result.

Currently, we are using a BLOSUM matrix to model the similarity between amino acids. However, such representation does not consider the neighboring amino acids, and lose local context. At this point, we are still working on a new way to add the local context when comparing proteins.

For our next step, we would like to test the model on a more diverse data set in order to discover protein structure patterns that are finer than the domain pattern.

V. Acknowledgement

This project was supported by the Jerome A. Schiff Fellowship, and I thank Prof. Pengyu Hong for his continuous support, inspiration and guidance.