A novel Greedy approach for Sequence based Computational prediction of Binding-Sites in Protein-Protein Interaction

***Abstract*—Computational prediction of protein-protein interaction (PPI) from protein sequence is important as many cellular functions are made possible through PPI. The Protein Interaction Prediction Engine (PIPE) software suite was developed at *Carleton University* for such predictions. The location of interaction is predicted by a followup PIPE-Site predictor, that depneds on PIPE engine. This PIPE-Sites predictor developed a decade ago and needed to be updated through the use of a large high-quality set of known PPI sites. Additionally, a similarity-weighted score had been recently developed in PIPE and has been proven to be more accurate for PPI prediction probability. However, PIPE-Sites are shown to be ineffective when applied to similarity-weighted score data. Thus, a new sequence-based method, named Panorama, of predicting PPI sites is proposed and evaluated in this. This new methods leverages similarity-weighted score data to further increase performance over two different performance metrics when evaluated on both human and yeast PPI site data.**

***Index Terms*—protein-protein interaction, Greedy Algorithm, Sequence based Protein-Protein Interaction**

I. Introduction

Proteins are bio-molecules that comprise the majority of the cellular machinery in any biological system. Proteins carry out many cellular functions through physical protein-protein interactions (PPI) with other proteins [1]. Such PPI are essential to many cellular processes including the regulation of biochemical pathways, cellular motion, forming protein complexes, or carrying another protein. Additionally, understanding how proteins interact, and evaluating the site of interaction for each protein, will provide information on complex networks, evolution, and human pathology [2].

The significance of understanding these PPIs led to the development of different experimental methods to detect and characterize such interactions including both experimental (in vivo or in vitro) and computational (in silico) strategies [3]. Experimental approaches, such as yeast two-hybrid and affinity purification followed by mass spectroscopy, tend to be expensive, labor intensive, time consuming and suffer from noise [4]. Therefore, computational techniques have been successfully used to predict high-throughput protein interaction data with high quality and accuracy [5]. Hence, using computational approaches for the prediction for PPIs are an effective alternative, particularly those that predict PPI based solely on the primary sequence of the two query proteins since three-dimensional protein structure is much more difficult to discern [6].

Beyond the prediction of whether two proteins will interact, characterization of the amino acid sub-sequences that support the PPI provides important insights [7]. These so-called PPI "interaction sites" include all amino acids that form the physical interaction interface and also the surrounding amino acids that are required to support the PPI (e.g., surrounding amino acids may form the structural scaffolding for a binding pocket). Interaction sites can be predicted from frequently occurring polypeptide sequences. The underlying principle of sequence-based predictors is the conservation of sub-sequences (often referred to as domains and motifs) that appear in many PPI sites. That is, if a subsequence is known to enable a PPI in one or more proteins, and that same subsequence is observed in the query protein, this suggests that the query protein may also participate in PPI.

The protein-interaction prediction (PIPE) is a high throughput, sequence-based method developed at Carleton University by the Bioinformatics Research Group [8]. This algorithm only requires the amino acid sequences of the proteins of interest, and a database of known PPIs. It uses a machine learning model to assign a probability of whether a pair of proteins will interact or not. In addition, the Protein Interaction Prediction Engine- sites (PIPE-Sites) software goes beyond PPI prediction to also determine which parts of each protein, or "PPI (binding) sites", are responsible for the interaction [7].

PIPE-Sites was also developed at Carleton University, but PIPE-Sites was trained using a relatively small dataset of gold standard PPI sites from yeast data, a decade ago. There are now much more PPI data available for different species. The PIPE-Sites method is parameterized by a number of tunable hyperparameters in its core algorithm: the "walk" algorithm. These hyperparameters were optimized and validated using only small datasets. Furthermore, recent improvements to the PIPE algorithm have accounted for frequently occurring subsequences that are not informative of protein interactions, leading to new similarity-weighted score-data (SW-score). The prediction of PPI interaction sites will likely benefit from leveraging these new SW-score data but may require retuning of the PIPE-sites hyperparameters or, indeed, replacement of the core algorithm.

References

[1] M. Kotlyar, C. Pastrello, A. E. Rossos, and I. Jurisica, "Protein-protein interaction databases," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 2018, vol. 1-3.

[2] F. H. Stephenson, *Calculations for Molecular Biology and Biotechnology: Third Edition*. Elsevier Inc., 2016.
[3] J. Zahiri, J. Bozorgmehr, and A. Masoudi-Nejad, "Computational Prediction of Protein–Protein Interaction Networks: Algorithms and Resources," *Current Genomics*, vol. 14, no. 6, 2013.
[4] M. Shatnawi, "Review of Recent Protein-Protein Interaction Techniques," in *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology: Algorithms and Software Tools*, 2015.
[5] T. Sun, B. Zhou, L. Lai, and J. Pei, "Sequence-based prediction of protein protein interaction using a deep-learning algorithm," *BMC Bioinformatics*, vol. 18, no. 1, 2017.
[6] K. Dick and J. R. Green, "Reciprocal Perspective for Improved Protein-Protein Interaction Prediction," *Scientific Reports*, vol. 8, no. 1, 2018.
[7] A. Amos-Binks, C. Patulea, S. Pitre, A. Schoenrock, Y. Gui, J. R. Green, A. Golshani, and F. Dehne, "Binding Site Prediction for Protein-Protein Interactions and Novel Motif Discovery using Re-occurring Polypeptide Sequences," *BMC Bioinformatics*, vol. 12, 2011.
[8] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Jessulat, N. Krogan, X. Luo, and A. Golshani, "PIPE: A protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs," *BMC Bioinformatics*, vol. 7, 2006.