Summary of Doctoral Dissertation
Protein-Protein Interaction Networks and Some Related Problems

PhD candidate: Nguyen Thanh Phuong Supervisor: Ho Tu Bao School: Knowledge Science

In all areas of biological and medical research, the role of the computer has been dramatically enhanced in the last five- to ten-year period. While the first wave of computational analysis did focus on sequence analysis, where many highly important unsolved problems remain. Outstandingly, protein-protein interaction research looks into the association of proteins to discover the rules controlling their interactions, which are key parts of cell mechanisms. In addition to the central problem of PPI prediction, two relevant problems have been raising and developing rapidly, i.e., the study of signal transduction networks (STN) and the study disease-causing genes. STN play an important role in the control of most fundamental cellular processes including cell proliferation, differentiation, and survival. It is known that STN most likely dependent on PPI. Discovering human disease-causing genes (disease genes in short) is one of the most challenging problems in bioinformatics and biomedicine, as most diseases are related in some way to our genes. Keeping with the most attractive problems, our study targets three significant problems: (1) protein-protein interaction prediction, (2) signal transduction construction, and (3) disease-causing gene prediction.

Since the experimental work itself much involves quantitative tasks, computer science came to the scene bringing about another approach, the computational one, to the standing issues. Computational methods become more and more essential to mine the huge amount of data and discovery useful knowledge for life science. In such context, our strategy is twofold. The first one is to take the full advantage of the biological nature of PPI, STN and disease-causing genes underlying a titanic amount of data. The second one is to develop appropriate and robust computational methods to integrate those complex biological and medical data, and then solve three targeted problems. These proposed methods fill the gaps of existing methods and achieve considerable contributions as follows.

1. *Protein-protein interaction prediction:* We proposed a novel integrative domain-based method to predict protein-protein interactions. The previous works used either multiple data sources as in integrative methods or only protein domain features as in the domain-based methods. The key idea of our computational method is to integrate protein domain features and genomic and proteomic features from multiple data sources into PPI prediction using Inductive Logic Programming (ILP). Comparing with other methods, our method outperformed in terms of several evaluation measures. Moreover, representing in forms of ILP rules, the predictions were easy to interpret and useful for biologists.
2. *Signal transduction construction***:** We developed an effective computational method to construct human signal transduction networks from protein-protein interaction networks. The proposed method was better than the previous ones by exploiting three biological facts of STN applied to human: (1) rich-information of protein-protein interaction networks, (2) signaling features and sharing components, and (3) then constructs STN effectively. We firstly consider different levels of signaling machinery in terms of various signaling features. Secondly, soft clustering well detected the sharing components among STN. The early work had been done for yeast, and later we shifted to human STN, a currently significant challenge. To the best of our knowledge, this study is the first one that has taken effort to construct human STN. Both the evaluation of STN construction for yeast and human are promising with high performance and gain some considerable findings.
3. *Disease-causing gene prediction:* We developed a new method for discovering disease genes with exploitation of semi-supervised learning, protein-protein interactions and multifarious biological features related to disease genes. Differed from existing work, our method based on semi-supervised learning (i) solves imbalance between known disease genes and unknown disease genes, (ii) integrates multifarious data related to disease genes, and (iii) exploit both useful information of labeled and unlabeled data. The contributions of this work are not only the new and effective method for disease gene prediction but also new significant findings. The comparative results demonstrated that our method obtained higher sensitivity, specificity, precision, accuracy, and balanced F-score. Testing with all interacting partners of disease proteins, we found 572 putative disease genes.

In conclusion, our effort in analyzing the interaction network data is to mine the coherent information, forecast unobserved interactions, and then detect relevant biological functions and processes, i.e., signal transduction networks and disease-causing genes. The thesis focuses on benefiting multiple data sources to solve three biologically significant problems in PPI research in both theoretical and empirical aspects. The theoretical aspect of this thesis concerns about the design of new and effective methods for protein-protein interaction prediction, signal transduction construction and disease-causing gene prediction. The other aspect is the application of these methods to produce lot of biological findings that can be useful sources for life scientists.