Chapter 7

Conclusions

7.1 Introduction

In the previous chapters of this thesis, we presented several approaches for advancing the state-of-the-art for a number of the tasks in the Protein-Protein Interaction (PPI) networks. In this final chapter, we discuss our main contributions and possible future trends for each open problem of PPI networks.

7.2 Shortest-Path Distance and Anova-based Feature Selection

We modeled the PPI network as a graph $G(V, E)$, where V is a set of nodes (proteins in our context) and E is a set of edges (interactions in our context) connecting pairs of nodes. Shortest-path distance is a simple and still powerful feature when the input data is modeled as a graph. In the context of PPI networks, this type of feature has been used for network clustering before. In chapters 2 and 4, we proposed to use this type of feature for predicting annotation information of proteins in the PPI networks. A general predicting procedure was as follows: First, we described the proteins based on their shortest-path distance to specific, automatically selected, other proteins in the PPI network. Second, we apply machine learning for the prediction task.

Noisy nature of PPI networks and high-dimensional description vectors in large graphs are potential problems of this general predicting procedure. We proposed to reduce the noise and dimensionality in the description vectors by only retaining the shortest-path distance to a few "important" nodes. We defined node i as an "important" node, if the shortest-path distance of some node v to i is likely to be relevant for v's classification. We applied the Anova measure to select the important proteins in the PPI networks. We used shortest-path distance as a predictive feature and the Anova measure as a feature selection strategy in chapters 2 and 4 of this thesis. In both cases, the empirical results proved the usefulness of the proposed features.

7.3 Collaborative Functions

One of the main open problems of PPI networks is to predict the functional annotation of proteins in the network. Most of the previous methods predict the proteins' functions based on guilt-by-association and here, we call it Similarity Assumption: Interacting proteins tend to have the similar functions. In chapter 3 of this thesis, we considered a biological process as an aggregation of each individual protein's functions. So, we assumed that topologically close proteins tend to have collaborative functions and not necessarily similar functions (Collaboration Assumption). We defined "collaborative functions" as pairs of functions that frequently interface with each other in different interacting proteins. To our knowledge, this was the first study that considered the collaboration assumption for the task of function prediction in

PPI networks. The information about which functions collaborate, can be extracted easily from a PPI network, and using that information leads to improved predictive accuracy. We proposed two methods for this purpose: The first method calculates the collaboration value of two functions based on an iterative reinforcement strategy. The second method adopts an artificial neural network. Empirical results confirmed that the notion of collaborativeness of functions, rather than similarity, is useful for the task of predicting the functions of proteins.

As a future works, we may apply this idea to other domains, outside PPI networks. The notion of homophily is well-known in network analysis; it states that similar nodes are more likely to be linked together. The notion of collaborativeness, in this context, could also be described as "selective heterophily". It remains to be seen to what extent this notion may lead to better predictive results in other types of networks.

7.4 Network Contextual Information

PPI networks have been widely used for the task of predicting proteins involved in cancer. When the input data is a PPI network, the main challenge is to find features with good predictive power that can be computed from this network. Previous machine learning based methods have mostly focused on the functional information about the protein for which a prediction is made, or proximity of known cancer-related genes in the PPI network. In chapter 4 of this thesis, we proposed the following two types of input features and we showed that these features have good predictive power.

1. Functional Context: While previous methods have considered GO annotations of proteins as predictive features, no methods up till now have considered GO annotations of the neighbors of those proteins at the same time. One advantage of using GO annotations of the neighbors for the prediction task is that GO annotations are often incomplete, and by collecting GO information from the neighbors of a protein p, we may get more information about p itself. This argument is backed up by the fact that GO annotations of proteins can often be predicted well from the GO annotations of their neighbors. However, this is not the only effect; there is also a direct relationship between a protein's involvement in cancer and the GO annotations of the proteins it interacts with.

2. Structural Context: This context relates to the relative position of proteins in the network. Several previous methods described each protein p based on the shortest-path distance of p to some previously known cancer/disease proteins. Alternatively, we could describe a protein's position relative to other proteins than only cancer-related ones. In this thesis, we proposed a relevance measure for proteins that is inspired by statistical Anova, and showed that shortest-path distance to a relatively small number of proteins (selected according to the Anova-based measure) is informative for the task of predicting cancer-related proteins in the PPI networks.

Empirical results proved that the proposed network contextual information (functional and structural contexts) of a protein in a PPI network, offer additional information regarding the possible involvement of a protein in cancer. These features increase the accuracy of predictive models and have a biological interpretation.

7.5 Interaction-based Chi-square

The task of predicting in a PPI network which proteins are involved in cancer has received a significant amount of attention in the literature. Several approaches have been proposed based on machine learning methods. Their success depends on two main parameters: First, feature representation of the proteins and second, choosing the right machine learning method. The previous methods studied these two parameters and found that the quality of the prediction results depends only slightly on the chosen machine learning method, but strongly on the chosen features, and after considering different protein's features individually, Gene Ontology (GO) annotations turned out to be particularly important. Several authors proposed to use a χ^2-based feature selection method to select the most relevant GO terms.

Selecting individual discriminative functions based on original χ^2 does not consider the network topology and the way different functions interact with each other in the network. For the task of predicting cancer-related proteins, it is possible that a function f_i does not correlate itself with cancer-involvement, but when a protein with function f_i interacts with a protein with function f_j, this interaction may be an indication of the former protein being involved in a cancer. In chapter 5 of this thesis, we proposed a new method, "Interaction-based Chi-square", to combine the GO annotations of proteins with the information contained in the topology of a PPI network for the feature selection task. Empirical results show that our proposed interactive features are biologically meaningful and improve the prediction accuracy of these systems.

7.6 Informative Human Disease Network

Identification of novel proteins likely involved in diseases is an important issue in the area of computational biology. Previous methods assumed to have a set of proteins which are previously known to be involved in disease (i.e., seed proteins) and then, they try to extend the seed proteins by predicting new disease-related proteins. In almost all the discussed methods, prediction accuracy depends directly on the initial seed proteins. While the initial seed proteins of each disease suffers from several 'False Negative' cases (i.e., disease-related proteins which are not annotated as being involved in disease), dependency of previous methods to the incomplete seed proteins is the main drawback of these methods.

In chapter 6 of this thesis, we reduced the number of False Negative cases in the initial seed proteins by proposing an informative Human Disease Network (HDN). We analyzed different *Structural* and *Functional* prediction methods and we concluded

that a hybrid method which considers both structural and functional information in the PPI network is the best method for building the HDN. We built a HDN based on 20 diseases and we showed that it is biologically meaningful. Then, we clustered the HDN and we augmented the seed proteins of diseases based on the cluster they belong to. Finally, we predicted disease-related proteins based on the augmented version of seed proteins. Literature mining of the newly found disease-related proteins proved the usefulness of our proposed HDN for predicting disease-related proteins.

7.7 Future Works

As a future works, we could apply our contributions to other domains, outside PPI networks. For example, Rahmani et al., [101] predicted the social tags in the graph of annotated web pages based on the "selective heterophily" notion discussed in chapter 3. They observed that this idea improves the prediction accuracy. We could also use our proposed HDN for a hypothesis generation about diseases' drug targets.