# Summary

The main goal of this thesis is mining annotated graphs. We chose Protein-Protein Interaction (PPI) networks as a specific graph domain to apply our methods. We modeled the PPI network as a graph where each node is a protein and each edge is a physical interaction between two proteins. There are different types of annotation information for each protein in the PPI network. "Functional annotation" states the biological functions of proteins in the PPI network and "disease/cancer relatedness annotation" indicates if one protein is involved in disease/cancer or not. We worked on these prediction tasks to improve the annotation information of proteins in the PPI network.

The task of function prediction in the PPI network is trying to predict the functions of un-annotated proteins based on the information in the network. We proposed two approaches for this task. In the first approach, we used shortest-path distances among different proteins as protein description features and Anova (Analysis of variance) as a feature selection method for reducing the noise and dimensionality in the description vectors. Then, we applied machine learning for the prediction task. In the second approach, we introduced novel functional features that indicate so-called "Collaborative Functions": Pairs of functions that frequently interface with each other in different interacting proteins. Most of the previous methods predict the proteins' functions based on guilt-by-association: Interacting proteins tend to have similar functions. We proposed two methods to extract collaborative functions from the PPI network. 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.

The task of predicting cancer-related proteins in the PPI network is trying to predict the new proteins involved in cancer. We generalized the previous methods as a two-steps algorithm. First, they select some features based on the training data to describe the proteins in the test data. Second, they apply machine learning methods to the description features in order to predict the new cancer-related proteins. Empirical results show that prediction accuracy depends more on the discriminative features rather than the machine learning methods and among different features ap-

plied individually, biological functions seems to be the most discriminative features. We proposed two approaches to select the novel features from the PPI network. In the first approach, we considered functional and structural contexts of proteins in the PPI network using the Anova measure and the chi-square method. In the second approach, we proposed a new method, "Interaction-based Chi-square", to combine the functional annotations of proteins with the information contained in the topology of a PPI network for the feature selection task. Empirical results showed that our proposed feature selection approaches are biologically meaningful and improve the prediction accuracy of these systems.

The task of predicting disease-related proteins in PPI network is an important issue in the area of computational biology. Previous methods assume 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. While the initial seed proteins of each disease is incomplete and suffers from 'False Negative' cases (i.e., disease-related proteins which are not annotated as being involved in disease), dependency of previous methods on the incomplete seed proteins is the main drawback of these methods. We proposed an informative Human Disease Network (HDN) considering both functional and structural information in the PPI network to reduce the number of False Negative cases in the initial seed proteins of 20 analyzed diseases. Literature mining of newly predicted proteins proved the usefulness of the proposed HDN.