Chapter IV
Incorporating Graph Features for Predicting Protein–Protein Interactions

Martin S. R. Paradesi
Kansas State University, USA

Doina Caragea
Kansas State University, USA

William H. Hsu
Kansas State University, USA

ABSTRACT

This chapter presents applications of machine learning to predicting protein-protein interactions (PPI) in Saccharomyces cerevisiae. Several supervised inductive learning methods have been developed that treat this task as a classification problem over candidate links in a PPI network – a graph whose nodes represent proteins and whose arcs represent interactions. Most such methods use feature extraction from protein sequences (e.g., amino acid composition) or associated with protein sequences directly (e.g., GO annotation). Others use relational and structural features extracted from the PPI network, along with the features related to the protein sequence. Topological features of nodes and node pairs can be extracted directly from the underlying graph. This chapter presents two approaches from the literature (Qi et al., 2006; Licamele & Getoor, 2006) that construct features on the basis of background knowledge, an approach that extracts purely topological graph features (Paradesi et al., 2007), and one that combines knowledge-based and topological features (Paradesi, 2008). Specific graph features that help in predicting protein interactions are reviewed. This study uses two previously published datasets (Chen & Liu, 2005; Qi et al., 2006) and a third dataset (Paradesi, 2008) that was created by combining and augmenting three existing PPI databases. The chapter includes a comparative study of the impact of each type of feature (topological, protein sequence-based, etc.) on the sensitivity and specificity of classifiers trained using specific types of features. The results indicate gains in the area under the sensitivity-specificity curve for certain algorithms when topological graph features are combined with other biological features such as protein sequence-based features.

INTRODUCTION

Protein-Protein Interaction Prediction Problem

The term *protein-protein interaction (PPI)* refers to associations between proteins as manifested through biochemical processes such as formation of structures, signal transduction, transport, and phosphorylation. PPI plays an important role in the study of biological processes. Many PPIs have been discovered over the years and several databases have been created to store the information about these interactions such as BIND (Bader *et al.*, 2003), DIP (Salwinski *et al.*, 2004), MIPS (Mewes *et al.*, 2002), IntAct (Kerrien *et al.*, 2007) and MINT (Chatraryamontri *et al.*, 2007). In particular, more than 80,000 interactions between yeast proteins are available from various high-throughput interaction detection methods (von Mering *et al.,* 2002). These methods can detect if the interaction is either a physical binding between proteins or a functional association between proteins. Often, the functional association between two proteins leads to physical binding among them. Determining PPI using high-throughput methods is expensive and time-consuming. Furthermore, a high number of false positives and false negatives can be generated. Therefore, there is a need for computational approaches that can help in the process of identifying real protein-protein interactions.

Several methods have been designed to address the task of predicting protein-protein interactions using machine learning. Most of them use features from protein sequences (e.g., amino acids composition) or associated with protein sequences directly (e.g., GO annotation). However, the PPI network can be used to design node and topological features from the associated graph. Several methods use such relational and structural features extracted from the PPI network, along with the features related to the protein sequence. This chapter provides an overview of several machine learning methods for predicting PPI using the graph information extracted from a PPI network along with other available biological features of the proteins and their interactions, and shows the importance of the graph features for accurate predictions.

Overview of PPI Databases

Several PPI databases have been used to extract examples of PPIs for machine learning algorithms. We review the main PPI databases in what follows.

The Biomolecular Interaction Network Database (BIND)

BIND (Bader *et al.*, 2003) stores information about interactions, complexes and pathways. It also contains a number of large scale interaction and complex mapping experiments using yeast two-hybrid, mass spectrometry, genetic interactions and phage display. The group that maintains BIND has also developed a graphical analysis tool that provides users an understanding of functional domains in protein interactions. They have also developed a clustering tool that allows users to divide the protein interaction network into specific regions of interest. BIND assumes that interactions can occur between two biological 'objects', which could be proteins, RNA or DNA sequences, genes, molecular complexes, small molecules, or photons (light).

The Database of Interacting Proteins (DIP)

DIP (Salwinski *et al.*, 2004) is a database containing 18,343 interactions between 4,923 proteins validated from 23,366 experiments of the *Saccharomyces cerevisiae* organism. A few of the experiments from which they validate protein interactions are co-immunoprecipitation, yeast two-hybrid and in vitro binding assays. The group that maintains DIP has developed several quality assessment methods and uses them to identify the most reliable subset of the interactions that are inferred from high-throughput experiments. They also provide an online implementation of their evaluation methods that can be used to evaluate the reliability of new experimental and predicted interactions.