Chapter IX
Discovering Protein Complexes in Protein Interaction Networks

Clara Pizzuti
ICAR, Consiglio Nazionale delle Ricerche, Italy

Simona Ester Rombo
Università della Calabria, Italy

ABSTRACT

In this chapter a survey on the main graph-based clustering techniques proposed in the literature to mine protein-protein interaction networks (PINs) is presented. The detection of putative protein complexes is an important research problem in systems biology. In fact it may help in understanding the mechanisms regulating cell life, in deriving conservations across species, in predicting the biological functions of uncharacterized proteins, and, more importantly, for therapeutic purposes. Different kind of approaches are described and classified. Furthermore, some validation techniques commonly exploited in this context are illustrated. The goal of the chapter is to provide a useful guide and reference for both computer scientists and biologists. Computer scientists may have a complete vision of what has already been made and which are the new challenges about PINs clustering, taking them as a starting point for further researches and new proposals; on the other hand, biologists may find in the chapter the necessary material to select the most appropriate methods to apply for their specific purposes.

INTRODUCTION

In the last few years the development of new high-throughput technologies (von Mering, 2002) to determine protein interactions has made available large volumes of experimental data that reflect the interplay between proteins in complex cellular networks. An important aspect of the protein-protein interaction networks (*PINs*) analysis concerns the discovery of *modules*, or *complexes*, made of proteins characterized by dense sets of interactions. The interest in this kind of analysis is motivated by the observation that PINs have modular structure (Hartwell, 1999; Pereira-Leal, 2004), being organized into different putative protein complexes each performing specific tasks in the cell. It is worth to point out that "protein complexes" and "function modules" have different biological meanings. A protein complex is a molecular machine that consists of several proteins that bind each

other at the same place and time. On the contrary, a functional module consists of a few proteins that control or perform a particular cellular function through interactions between themselves (these proteins do not necessarily interact at the same time and place). However, in many cases it is hard to distinguish these two concepts because the analyzed pair-wise protein interactions are not associated with temporal and spatial information. Thus, in the following we will refer indifferently to protein complexes or functional modules.

The detection of putative protein complexes may help in understanding the mechanisms regulating cell life, in deriving conservations across species, in predicting the biological functions of uncharacterized proteins, and, more importantly, for therapeutic purposes. Indeed, it has been recognized by biologists that proteins interacting with each other often participate in the same biological processes. Furthermore, protein modules may be often associated with specific biological functions and proteins belonging to a specific module are more related each other than w.r.t. the members of other modules (Tornow, 2003). Thus, discovering putative protein complexes may be useful also to detect processes in which uncharacterized proteins are involved. Unfortunately, isolating protein complexes through the PINs exploration is per se a difficult problem, due to the intricateness of interaction relationships characterizing them (see Figure 1, where only a small part of *S. cerevisiae* protein interactions are shown).

Recently, clustering techniques (Jain, 1988) proved to be successful for the identification of putative protein complexes in complex networks. Clustering consists in grouping data objects into groups (clusters) such that the objects in the same cluster are more similar each other than with objects in the other clusters. In PINs, clustering means grouping together proteins which share a larger number of interactions, thus clusters may be assimilated to functional modules. Possible uncharacterized proteins in a cluster may be assigned to the biological function recognized for that module. A number of clustering approaches have been proposed to extract relevant modules from PINs. Some of them rely on distance-based hierarchical clustering methods (Blatt, 1996; Arnau, 2004; Pei, 2005). In these approaches suitable similarity metrics have to be defined. Others consider the topology of the PIN and apply different strategies such as searching for sub-graphs having maximum density (Bader, 2003; Derenyi, 2005; Palla, 2005; Adamcsek, 2006; Altaf, 2006; Lubovac, 2006; Chua, 2007; Li, 2007; Pizzuti, 2007; Pizzuti, 2008); optimization of a cost function (King, 2004; Spirin, 2003); flow simulation (Enright, 2002; Pereira, 2004; Cho, 2006; Hwang, 2006; Cho, 2007); statistical approaches (Samantha, 2003; Farutin, 2006). Some technical and biological distinction among such approaches may be driven by considering if they possibly admit clustering overlapping, i.e., they allow proteins in the same network to participate in several clusters. In fact, as will be underlined in the following sections, there exist proteins which are involved in several biological processes, thus, methods that output overlapping clusters are biologically more relevant.

Figure 1. Some of the S. Cerevisiae protein interactions