Chapter 43
Using a Genetic Algorithm and Markov Clustering on Protein-Protein Interaction Graphs

Charalampos Moschopoulos
Biomedical Research Foundation of the Academy of Athens, Greece & University of Patras, Greece

Grigorios Beligiannis
University of Western Greece, Greece

Spiridon D. Likothanassis
University of Patras, Greece

Sophia Kossida
Biomedical Research Foundation of the Academy of Athens, Greece

ABSTRACT

In this paper, a Genetic Algorithm is applied on the filter of the Enhanced Markov Clustering algorithm to optimize the selection of clusters having a high probability to represent protein complexes. The filter was applied on the results (obtained by experiments made on five different yeast datasets) of three different algorithms known for their efficiency on protein complex detection through protein interaction graphs. The results are compared with three popular clustering algorithms, proving the efficiency of the proposed method according to metrics such as successful prediction rate and geometrical accuracy.

INTRODUCTION

The importance of protein interactions is given as they play an important role on fundamental cell functions. They are crucial for forming structural complexes, for extra-cellular signaling, for intracellular signaling (Ryan & Matthews, 2005). Recently, new high throughput experimental methods (Ito et al., 2001; Puig et al., 2001; Stoll, Templin, Bachmann, & Joos, 2005; Willats, 2002) have been developed which detect thousands protein-protein interactions (PPIs) with a single experiment. As a result, enormous datasets have been generated which could possibly describe the functional organization of the proteome. However, these data are extremely noisy (Sprinzak, Sattath, &

DOI: 10.4018/978-1-4666-3604-0.ch043

Margalit, 2003), making it difficult for researchers to analyze them and extract valuable conclusion such as protein complex detection or characterizing the functionality of unknown proteins.

Due to the vast volume of PPI data, they are usually modeled as graphs G=(V,E) where V is the set of vertices (proteins) and E the set of adjacent edges between two nodes (protein interactions). The model of graph makes it easy for bioinformatics researchers to apply various algorithms derived from graph theory in order to perform clustering and detect protein complexes which are represented as dense subgraphs (Bader & Hogue, 2003; Hartuv & Shamir, 2000; Koyuturk, Szpankowski, & Grama, 2007). According to Broheeand van Helden (2006) and Li, Wu, Kwoh, and Ng (2009), the most prevailed algorithms are MCL (Markov clustering) (Enright, Van Dongen, & Ouzounis, 2002) and RNSC (Restricted Neighbourhood Search Clustering) (King, Przulj, & Jurisica, 2004). Besides them, spectral clustering can achieve similar results (Kritikos, Moschopoulos, Vazirgiannis, & Kossida, 2011). While these methods use the PPI graph structure to detect protein complexes, additional information could also be used such as gene expression data (Eisen, Spellman, Brown, & Botstein, 1998), functional information (Gene Ontology Consortium, 2006) as well as other biological information (Huh et al., 2003). However, the use of additional information has the disadvantage that cannot cover the aggregation of proteins that constitute the PPI graph.

The aforementioned algorithms assign each protein of the initial PPI graph to a cluster, constructing clusters that could hardly be characterized as dense ones. As a result, their prediction rate of protein complexes is pretty low. One way to deal with this problem is to filter the results of such an algorithm using additional information such as Gene Ontology (King et al., 2004). However, the sources of the additional information usually do not cover all the recorded interactions that form the PPI graphs. Moreover, the parameters of these filters are almost always empirically defined, leading to biased solutions.

In this contribution, a filter is constructed by four methods which are based on graph properties such as density, haircut operation, best neighbor and cutting edge and it is applied on the results of MCL, RNSC and spectral algorithm. Furthermore, the parameters of the filter methods are optimized by a Genetic Algorithm (GA) which takes into account the rate of successful prediction, the absolute number of valid predicted protein complexes and the geometrical accuracy of the final clusters. Extended experiments were performed using five different PPI datasets. The derived results were compared with the recorded protein complexes of the MIPS database (Mewes et al., 2006), while statistical metrics were calculated such as sensitivity (Sn), positive predictive value (PPV) and geometrical accuracy (Acc_g). To demonstrate the efficiency of the proposed filter, we compare the derived results with 3 other algorithms; SideS (Koyuturk et al., 2007), Mcode (Bader & Hogue, 2003), and HCS (Hartuv & Shamir, 2000).

The remaining of the paper is organized as follows: in the next section, clustering algorithms that have been applied on protein complex detection through PPI graphs are presented. In the third section, a thorough description of genetic algorithms and the main applications of them in Bioinformatics are given. In the forth section, our proposed methodology is presented while in fifth section we present and discuss the results of our experiments, carried out on five datasets (derived by databases or individual experiments) and by comparing our method with three other algorithms (Mcode, HCS, SideS). Finally, in the conclusion, we conclude and the main directions of future work are suggested.

RELATED WORK

In this section the functionality of the tested algorithms is analysed.