Multi-objective optimization for relevant protein complex extraction

Mohamed Elati, Cuong To and Rémy Nicolle

Institute of System and Synthetic Biology, University of Évry, 5 rue Henri Desbruères, 91030 Evry Cedex, France.
{cuong.to, remy.nicolle, mohamed.elati}@issb.genopole.fr

Introduction:

Protein complexes play central roles in many cellular pathways. Although many high-throughput experimental techniques such as yeast two hybrid have already enabled systematic screening of pairwise protein-protein interactions en masse, the amount of experimentally determined protein complex data has remained relatively lacking.
In a protein interaction graph, these complexes should be represented by a dense subgraph between the proteins forming a complex. As such, researchers have begun to exploit the large-scale protein-protein interaction data to help discover new protein complexes using graph clustering or dense subgraph extraction techniques. However, reliability of extracted protein complexes in PPI networks is not satisfactory because there are many data artifacts in the underlying protein-protein interaction data due to the limitations in the high-throughput screening methods.
At the same time, affinity purification followed by mass spectrometry analysis (AP-MS) is a context-specific technique used to identify the protein partners (hit proteins) of a particular protein of interest (bait protein). AP-MS has been used to identify proteins that are members of the same protein complex as the bait. There are some specific challenges associated with MS data, false positives and missed proteins can occur. In particular, false positives can arise when proteins are quantitatively abundant. Next, AP-MS data does not provide direct measurement of protein-protein interactions, but rather a set of proteins that are pulled together by a bait.

Results:

We introduce a new methodology to amalgamate the information from context-specific MS data with large-scale protein-protein interaction data.

We propose a novel algorithm that considers extracting relevant sub-graphs connecting pre-defined protein sets (MS data) from PPI networks as multi-objective optimization, density of the sub-graph and number of pre-defined proteins. We then exerted Genetic Algorithms (GAs) to solve the multi-object optimization. Our algorithm also introduced a novel solution representation called circle where the dimension of the search space is fixed to two dimensions, the center and the radius. Usually a solution is described as a set of proteins for which the dimension is rather high and not fixed. The application of this algorithm to real AP-MS data (Gavin *et. al., 2006)* and protein interaction database BIOGRID shows its ability to find dense and relevant protein complexes in large biological networks and outperforms existing method like MCODE (Bader and Hogue, 2003) and MCL (Van Dongen, S, 2000). We also show that from a potential protein complex, our algorithm is able to add only a few proteins, which display distinctive network topological features and molecular function annotations, and can be proposed as putative new components.

References

1. Van Dongen, S.: Graph clustering by flow simulation. PhD Thesis, University of Utrecht, Utrecht, The Netherlands (2000)
2. Bader, G.D., Hogue, C.WV.: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 4:2 (2003)
3. Gavin, A.C., *et al*.: Proteome survey reveals modularity of the yeast cell machinery. Nature, 440, 631-636 (2006