A Framework for the Prediction of Protein Complexes

ID - 158

Cannataro Mario[1], Pietro Hiram Guzzi[1], Pierangelo Veltri[1]

[1]Laboratory of Bioinformatics, University Magna Graecia of Catanzaro, Catanzaro, Italy

Motivation

Cellular processes are composed by a set of elementary events mediated by proteins.

Identifying and analyzing all the possible physical interactions among proteins do appear as an important step in studying cellular biology.

Protein interactions are modeled as undirected graphs where vertices represent proteins and edges denote interactions among them. Resulting networks of interactions are thus studied by investigating graph properties, such as connectivity or degree distribution, or by individuating particular regions that encode relevant biological properties. Searching for small and highly interconnected regions is used for the prediction of protein complexes [1], i.e. a set of proteins assembled together playing a biological function. Currently, there exist different approaches for the prediction of complexes, that are based on a particular graph clustering schema [1, 2, 3]. The in silico prediction of possible complexes could drive the planning of in vitro investigations avoiding a set of redundant or negative experiments.

However the scenario above described presents two main drawbacks: (i) currently available predictors may be only used by experts, (ii) results of different algorithms can not easily be integrated. In this work we propose a meta-predictor based on a novel algorithm that integrates results produced by different predictors to improve the prediction of complexes, and offers a graphical user interface to simplify the use.

Methods

Let us consider a set of pairwise interaction data publicly available on different datasets, e.g. MINT (http://cbm.bio.uniroma2.it/mint/) or MIPS (http://mips.gsf.de/genre/proj/mpact), merged in a comprehensive graph, and a set of n predictors that generate n different predictions. Each of these predictions is composed, obviously, by a set of different sub graphs, each one representing a possible protein complex.

The proposed algorithm builds a metaclustering in three steps: (i) in the first step it considers the graphs commonly found by the different predictors, and then it finds both (ii) sub/supergraphs (ii), and (iii) overlapping graphs.

The metaclustering algorithm is embedded into a more general framework organized in different modules: (i) a metaclustering module, the core of the architecture, that implements the algorithm, (ii) a set of clustering modules that wrap the existing clustering tools, (iii) a visualizer module that visualizes networks and results, and a (iv) data manager module that manages different formats of data and stores a reference dataset providing the evaluation of results.

Results

Currently we designed the overall architecture of the framework and we implemented the metaclustering and the data manager modules that provide the core functionalities of the framework.

In a first experiment, we considered a dataset of yeast described in [4] and available at http://rsat.scmbb.ulb.ac.be/sylvain/clustering_evaluation/. It contains 1095 nodes and 14658 vertices and it has been obtained by considering first the real interactions stored in the MIPS database and then by adding and removing randomly several edges to such data, simulating noisy data [4].

We run the MCODE [1], RNSC [2] and MCL [3] clustering algorithms giving such dataset as input. Then the obtained cluster data were used by the proposed algorithm to build different metaclusters, by considering different parameters. For each experiment we measured the Sensitivity, the Positive Predictive Value, and the Accuracy of metaclustering, as defined in [4]. The first one measures the fraction of proteins of a complex found in a common cluster, the second one represents the ability of a cluster to predicts a complex, while the last one is a geometric average. Results confirm that the metaclustering outperforms the other algorithms with respect to sensitivity and accuracy, by respectively 0,76 and 0,67 against 0,47 and 0,57 of the better algorithm, while they show a slight decrease of PPV.

In summary, the proposed meta-predictor enhances sensitivity and accuracy of the complex prediction with respect to the used predictors. More information are available at http://poseidon.bioingegneria.unicz.it/guzzi/projects.html.

References:

[1] Gary Bader et al, An automated method for finding molecular complexes in large protein interaction networks, BMC Bioinformatics 4 (2003), no. 1, 2.

[2] King A.D., Graph clustering with restricted neighbourhood search, Master thesis, University of Toronto, Toronto, Ontario, 2004.

[3] Van Dongen S et al, An efficient algorithm for large-scale detection of protein families, Nucleic Acids Research 30 (2002), no. 7, 1575_1584.
[4] Brohee et al, Evaluation of clustering algorithms for protein-protein interaction networks,BMC Bioinformatics 2006, no 7:488.

Availability: http://poseidon.bioingegneria.unicz.it/˜guzzi/projects.html
Email: cannataro@unicz.it