An efficient supervised method to integrate multiple biological networks

Alberto Bertoni, Marco Frasca, Giorgio Valentini
DSI – Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano.

Motivation: Biological networks can represent different types of relationships between biological entities (e.g. genes or proteins), ranging from genetic or physical interactions, to gene expression correlations or co-occurences in bio-medical literature.

In this context, a central problem is the integration of different networks in order to infer the underlying biological properties of the biological entities, i.e. the functional classes of genes, or the potential protein targets of a drug, with relevant applications in functional genomics, proteomics and pharmacogenomics.

Different approaches have been proposed in the literature, including conjunctive/disjunctive techniques, weighted combination of multiple sources of data and network integration through constrained linear regression techniques.

Despite these efforts, these methods suffer of drawbacks and limitations: they can be effective but their computational complexity can be prohibitive when very large biological networks are involved; on the other hand other methods can be computationally efficient but at the expenses of a coherent and effective integration.

Methods: We propose a novel supervised method to efficiently integrate multiple biological networks by combining edge weights according to the estimated accuracy achievable on each available data source.

By this approach relational data are transformed into two-dimensional vectors: each node, representing a biological entity of the network, becomes an instance to be classified. By applying an efficient linear classifier to the transformed two-dimensional data we obtain a measure of the linear separability of the nodes labeled according to the biological property under study. This measure can be used to weight different biological networks and the corresponding adjacency matrices can be easily integrated using the estimated weights. The method can also be applied to estimate whether a given source of relational data can be informative to predict a specific biological property. The algorithm is fast, scales nicely when new sources of data are added, and can be efficiently applied to large biological networks.

Results: We applied the proposed method to the integration of multiple sources of biomolecular data to predict the functional classes of genes in the yeast and mouse model organisms at genome-wide level, using the FunCat and Gene Ontology taxonomies. More precisely, we integrated protein domain, gene expression, protein interactions, and other types of biomolecular data using our proposed approach. A label propagation method based on cost-sensitive Hopfield networks has been used on both single and integrated data sets to infer the functional classes of genes.

Fig. 1 shows a summary of the ontology and genome-wide predictions of FunCat classes in the yeast, in terms of the F-scores averaged across the functional classes. Crossvalidated results show that our integration method significantly improves sensitivity and precision with respect to the best results achieved with single sources of data. Moreover F-score results are competitive with those obtained by other state-of-the-art supervised and semi-supervised methods proposed in the literature.

Figure1. *S. cerevisiae*: average precision, sensitivity, F-score across 168 FunCat classes, using single data sources and data integrated through our proposed method. Pfam1 and Pfam2 refers to protein domain data obtained from the Pfam data base; Expr to gene expression data from Spellman and Gasch experiments; PPI-BG and PPI-VM to protein-protein interaction data obtained respectively from the BioGRID and STRING databases.

Single data sources			
Dataset	*Precision*	*Sensitivity*	*F-score*
Pfam1	0,448	0,371	0,381
Pfam2	0,400	0,348	0,316
Expr	0,107	0,586	0,135
PPI-BG	0,500	0,478	0,465
PPI-VM	0,436	0,433	0,417
Integrated data			
Dataset	*Precision*	*Sensitivity*	*F-score*
Integrated data	0,562	0,490	0,506