# Data fusion based gene function prediction using ensemble methods

Re M[(1)], Valentini G[(1)]

[(1)] Computer Science Department (DSI), University of Milan, Milan

## Motivation

The integration of multiple sources of heterogeneous biomolecular data is a key item for the prediction of gene functions at genome-wide level. Several approaches have been proposed in the literature, ranging from function linkage networks, to graphical models, vector-space integration and kernel fusion methods. Nevertheless, all these approaches suffer of limitations and drawbacks, due to their limited scalability to multiple data, to their limited modularity when new data sources are added, or when the available biomolecular data are characterized by different structural features. A new possible approach is based on ensemble methods, i.e. committees of learning machines, but to our knowledge not too much works have been proposed in literature. There are several reasons to apply ensemble methods in the context of genomic data fusion for gene function prediction. At first, biomolecular data that differ for their structural characteristics (e.g. numerical vectors, strings or graphs) can be easily integrated, because with ensemble methods the integration is performed at the decision level. Moreover as new types of biomolecular data are made available, previously trained ensembles are able to embed new data sources by training only the base learners devoted to the newly added data, without retraining the entire ensemble. Finally most ensemble methods scale well with the number of the available data sources, and problems affecting other data fusion approaches are thus avoided.

## Methods

We performed our tests by integrating S.cerevisiae data collected from literature and comprising protein-protein interactions, protein domains, expression data and BLASTP pairwise similarities. The investigated genes were labelled according to the functional annotations available in the MIPS Functional Catalogue (FunCat) version 2.1 considering only the 15 functional classes at highest level in the FunCat hierarchy. The gene function prediction has been performed in a binary classification setting classifying all the genes as belonging to the current “target” functional class or to “other functional classes”. We applied a sigmoid fitting to the output of SVMs (each trained on different datasets), in order to obtain an estimate of the probability that a given example belongs to a functional class. We then combined the output produced by the SVMs through the classical weighted average rule, using weights calculated according to a convex combination rule and a logarithmic transformation, and the Decision Templates combiner. According to the test and select methodology, we applied a variant of the "choose the best" technique to select, for each function prediction task, several subsets of "optimal" classifiers. The investigated ensemble methods were then used to combine the outputs produced by all the component classifiers and all the selected combinations of base learners. Then we added a simple feature selection method based on the t-test and Benjamini-Hochberg p-value correction to select the most relevant features and to reduce the high dimensionality that characterize biomolecular data.

## Results

We compared the performances obtained by the tested strategies under different experimental conditions in order to provide an overview of the capabilities of ensemble systems in data fusion mediated gene function prediction. The ensembles outperformed the base learners in all the function prediction tasks. Performances are also improved by the applied classifier selection strategy and the feature filtering method. Considering the F-measure that summarizes both precision and recall, the experimental results show that data fusion realized by means of ensemble systems is a valuable research line in gene function prediction and that Decision Templates may represent a good choice for biomolecular data integration (see Figure).

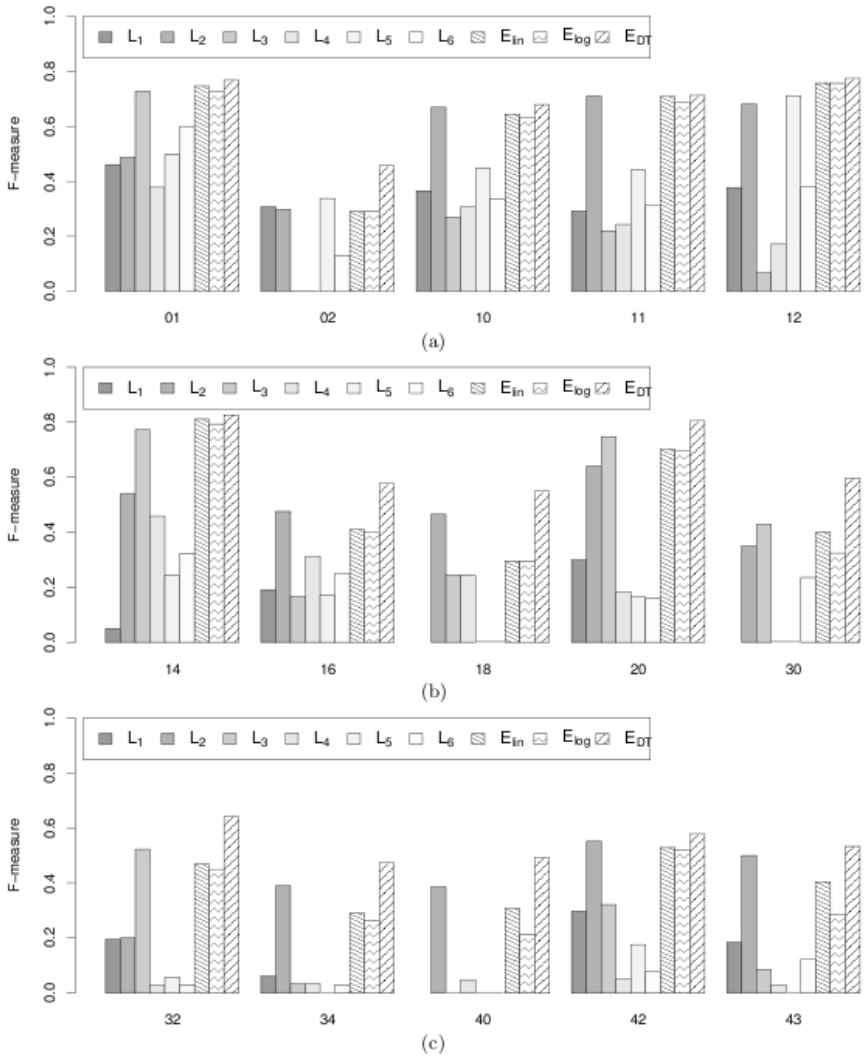


**Fig. 1.** Per class F-measure results of ensemble methods with base learner selection and feature filtering. For each FunCat class, the first six shaded grey bars refer to single learners with feature filtering (from $L_1$ to $L_6$); the last three bars (filled with patterns) correspond respectively to weighted linear combination with linear ($E_{lin}$) and logarithmic ($E_{log}$) weights and decision template ($E_{DT}$) ensembles. a) Funcat classes 01, 02, 10, 11, 12; b) 14, 16, 18, 20, 30; c) 32, 34, 40, 42, 43.

**Contact :** {re,valentini}@dsi.unimi.it