Prediction of gene function using ensembles of SVMs and heterogeneous data sources

Matteo Re and Giorgio Valentini

Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Italy
email: {re,valentini}@dsi.unimi.it

Summary. The ever increasing amount of biomolecular data available in public domain databases for a broad range of organisms coupled with recent advances in machine learning research has stimulated interest in computational approaches on gene function prediction. In this context data integration from heterogeneous biomolecular data sources plays a key role. In this contribution we test the performance of several ensembles of SVM classifiers, in which each component learner has been trained on different types of data, and then combined using different aggregation techniques. The compared combination methods are the widely adopted linear weighted combination, the logarithmic weighted combination and the similarity based decision templates approach. The results show that heterogeneous data integration through ensemble methods represents a valuable research line in gene function prediction.

1 Introduction

Functional classification of unannotated genes and the improvement of the existing gene functional annotation catalogs, are of capital importance in modern functional genomics and bioinformatics. Gene functional classification may provide useful insights in pharmacogenomics, being able to provide indications for the development of target specific drugs. More in general, it plays a key role in molecular biology, given it's ability to detect previously unknown role of genes and their products in physiological and pathological processes. Nevertheless, the application of automated systems in this research area is strongly limited by the intrinsic difficulty of this task, which is mainly originated by the natural heterogeneity of the involved data. Different types of biomolecular data, ranging from expression profiles to phylogenetic gene-specific evolution rates and many others can in principle provide useful information for the automated assessment of the functional role of genes. The extent of the degree at which the presence of a specific type of experimental data could result into an improvement of the classification performances is expected to vary ac-

cordingly to the specific gene and the particular bio molecular process under investigation.

Several approaches for heterogeneous biomolecular data integration have been proposed in the literature. A first one corresponds to the "early integration" technique, by which different vectorial data are concatenated [1]. Other methods are based on modeling networks of functional relationships between proteins; in this context graphical models provide a probabilistic framework for data integration [2]. Kernel methods and techniques based on kernel fusion methods represent another important research area with significant applications in the integration of different bio-molecular data sources for gene function prediction [3].

In the aforementioned scenario the application of methods able to deal with both different data sources and the problem to integrate the prediction obtained from different learners is clearly appealing. It is widely accepted that combining multiple classifiers can provide advantages over the monolithic approach to pattern classifier design [4], but a systematic evaluation of the impact on classification performances of different combination rules suitable to merge the output of gene function classifiers trained on different data sources, as today, has not been explored. To our knowledge, only some works have been proposed, such as the "late integration" of kernels trained on different sources of data [1], or the Naive-Bayes integration of the outputs of SVMs in the context of the hierarchical classification of genes [5]. In this work we investigate the effectiveness of three classifier fusion strategies using ensembles of Support Vector Machines [6] each of which trained to produce a probabilistic-like classification output [7].

In the next section we present the ensemble methods we used in our experiments. In Sect. 3 we describe the different types of high-throughput biomolecular data and the experimental setting we adopted to classify yeast genes according the highest level classes of the FunCat taxonomy [8]. Sect. 4 presents the main results obtained by comparing performances of single SVMs with respect to to ensembles that merge 6 different sources of biomolecular data. The conclusions summarize the main achievements and drawbacks of the proposed data fusion ensemble approach.

2 Methods

Ensembles of classifiers have enjoyed great attention because of their excellent generalization performances on a wide spectrum of applications. One of the main ideas behind the effectiveness of ensemble systems is that if the single classifiers composing the ensemble are diverse, then they are expected to make different errors, and combining the output produced by these classifiers can in principle reduce the error through averaging [9].

Diversity can be achieved using different sources data, thus obtaining different "views" of the same phenomenon. In particular the objective of data

fusion is to extract complementary pieces of information from different data sources and then merge them achieving a more informed decision about the phenomenon under analysis. Working with heterogeneous data sources, data fusion can be realized by means of an ensemble system composed by learners trained on different datasets and then combining the outputs of the component learners.

The continuous output assigned to an instance vector $\mathbf{x}$ by a binary classifier can be interpreted as the support given to the membership of $\mathbf{x}$ to a specific functional class. In particular, with SVMs, a probabilistic output can be obtained by applying a sigmoid fitting to their output [7]. As a consequence a trained classifier computes a function $d_j : X \rightarrow [0, 1]$ that estimates the probability that a given example $\mathbf{x} \in X$ belongs to a specific class ω_j. An ensemble combines the outputs of T base learners using a suitable combining function g to compute the overall support (e.g. the probability) μ_j for a given class ω_j:

$$\mu_j(\mathbf{x}) = g(d_{1,j}(\mathbf{x}), \ldots, d_{T,j}(\mathbf{x})) \tag{1}$$

2.1 Linear weighted combination with linear and logarithmic weights

Among the algebraic combiners, the simplest is the mean rule, which calculates the support μ_j for the membership of a current instance $\mathbf{x}$ to the ω_j class as the average of all classifiers outputs:

$$\mu_j(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} d_{t,j}(\mathbf{x}) \tag{2}$$

In our experiments we used the weighted average rule, in order to take into account the reliability of each base learner in the computation of the support μ_j:

$$\mu_j(\mathbf{x}) = \sum_{t=1}^{T} w_t d_{t,j}(\mathbf{x}) \tag{3}$$

The weights w_t are usually computed using an estimate of the overall accuracy of the base learners, but in our experimental setting, where the gene functional classes are largely unbalanced (positive examples are largely less than negative ones), we chose the F-measure (the harmonic mean between precision and recall) to compute the weights:

$$w_t = \frac{F_t}{\sum_{t=1}^{T} F_t} \tag{4}$$

The F-measure F_t of the t^{th} base learner can be estimated by "internal" cross-validation on the training set.

It can be shown [10] that if we have T independent classifiers each of which associated with some performance measure (such as the accuracy or the F-measure), the accuracy of an ensemble produced by combining the learners outputs by weighted majority voting is maximized if the output weights satisfy the proportionality

$$w_t \propto log\frac{p_t}{1-p_t} \tag{5}$$

where p_t is an estimate of the reliability of the t^{th} base learner (e.g. accuracy or F-measure).

In our experiments we implemented the weighted logarithmic combination by adding a small ε in order to avoid division by zero in eq. 5, and then by normalizing in order to obtain positive weights that sum to 1:

$$\hat{w}_t = ln\frac{F_t+\varepsilon}{1-F_t+\varepsilon} \tag{6}$$

$$w_t = \frac{\hat{w}_t - ln\varepsilon}{\sum_{t=1}^{T}(\hat{w}_t - ln\varepsilon)} \tag{7}$$

Once computed the weights for each classifier, according to eq. 4 for the linear weighted combination or to eq. 7 for logarithmic weighted combination, the final decision $D_j : X \rightarrow \{0,1\}$ of the ensemble is taken using the probability μ_j for the class ω_j (eq. 3):

$$D_j(\mathbf{x}) = \begin{cases} 1, & \text{if } \mu_j(\mathbf{x}) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where output 1 corresponds to positive and 0 to negative predictions for ω_j.

2.2 Decision Templates

The main idea behind decision templates is to compare a "prototypical answer" of the ensemble for the examples of a given class (the template), to the current answer (the decision profile) of the ensemble to a specific example whose class needs to be predicted [11].

The decision profile DP($\mathbf{x}$) for a instance $\mathbf{x}$ is a matrix composed by the $d_{t,j} \in$[0,1] elements representing the support given by the t^{th} classifier to class ω_j. The decision profiles matrices are effective tools that allows us to effectively summarize the information produced by all the members of an ensemble system and also provide conceptual blocks at the basis of the decision templates technique.

Decision templates DT_j are the averaged decision profiles obtained from $\mathbf{X}_j$, the set of training instances belonging to the class ω_j:

$$DT_j = \frac{1}{|\mathbf{X}_j|}\sum_{\mathbf{x}\in\mathbf{X}_j} DP(\mathbf{x}) \tag{9}$$

Note that the sum in eq. 9 refers to matrices, and hence decision templates are matrices with a number of rows equal to the number of the base learners and a number of columns equal to the number of the classes.

Given a test instance we first compute its decision profile and then we calculate the similarity S_j between $DP(\mathbf{x})$ and the decision template DT_j for each class ω_j. As similarity measure the Euclidean distance is usually applied:

$$S_j(\mathbf{x}) = 1 - \frac{1}{T \times C} \sum_{t=1}^{T} \sum_{k=1}^{C} [DT_j(t,k) - d_{t,k}(\mathbf{x})]^2 \tag{10}$$

The final decision of the ensemble is taken by assigning a test instance as:

$$D(\mathbf{x}) = \arg\max_j S_j(\mathbf{x}) \tag{11}$$

In our experimental setting we consider dichotomic problems, thus obtaining two-columns decision template matrices. Note that for each gene functional class we have two decision templates, one for the positive examples for that class (DT_P) and another one for the negatives (DT_N).

It is easy to see that with dichotomic problems the similarity measure (eq. 10) for the positive (S_P) and negative (S_N) class becomes:

$$S_P(\mathbf{x}) = 1 - \frac{1}{T} \sum_{t=1}^{T} [DT_P(t,1) - d_{t,1}(\mathbf{x})]^2 \tag{12}$$

$$S_N(\mathbf{x}) = 1 - \frac{1}{T} \sum_{t=1}^{T} [DT_N(t,1) - d_{t,1}(\mathbf{x})]^2 \tag{13}$$

and the final decision of the ensemble is:

$$D(\mathbf{x}) = \arg\max_{\{P,N\}} (S_P(\mathbf{x}), S_N(\mathbf{x})) \tag{14}$$

3 Experimental setup

3.1 Heterogeneous biomolecular datasets

In order to test the effectiveness of various continuous-valued predictions fusion methods we collected a set of data sources used in bioinformatics experiments published in literature or obtained from public databases. We chose to perform our experiments using data collected on *S. cerevisiae* because it is among the most studied and well characterized model organisms and because of the great amount of biomolecular data available for this species.

Biological functions are mediated in cell by several types of molecules but, in the large majority of biological processes, the final effectors are proteins,

Despite the complexity of single proteins the realization of a single biomolecular process usually requires the coordinated action of more than a single molecule, and the composition of the set of molecules preposed to the realization of the steps involved in the entire process is expected to be highly informative We thus decided to use protein-protein interaction data collected from BioGrid [12], a database of protein and genetic interactions and from STRING [13], a collection of protein functional interactions inferred from heterogeneous data sources, comprising, among the others, experimental data and information found in literature.

The evolutionary pressures acting during evolution on genes lead progressively to a saturation of the number of mutations that can be tolerated without disrupting the functionality of gene products and this reduces the ability to evolve new biomolecular functions. The conservative action of evolutionary pressures is particularly strong for single-copy genes but act in a more relaxed fashion on members of clusters of genes derived from duplication events in entire genomic regions. Provided the presence of multiple copies of a particular gene, the organism is allowed to explore evolutionary paths that would be otherwise precluded and that can lead to the evolution of novel functions by means of the changes occurring at nucleotide level in DNA sequences encoding protein products. As the evolutionary time increases, the genes belonging to the gene cluster will get even different at both nucleotide and aminoacidic sequence level, but the relations between members of the same gene families, which often share a similar biological role, can be detected using classical alignment approaches such as those proposed in [14] and [15]. The use of this type of experimental data enables the detection and quantification (using opportune similarity measures) of homology relationships occurring between genes through a simple nucleotide sequences comparison. In the aim to catch homology and, hopefully, functional relations existing between genes belonging to the same functional classes, we included as data source into our experimental framework the Pairwise Similarity Smith-Waterman dataset published in [3] (data kindly provided by the authors).

Even if protein-protein interactions and evolutionary signatures can provide useful information for functional classification of genes, a potential source of information about the functional role of a gene could be provided by the tight connection between the structure of a protein and its ability to perform a particular biological task. Proteins are constituted by structured regions usually referred as domains joined by unstructured regions named loops. Each specific domain constituting a protein is preposed to the realization of a specific task (either structural or biochemical) and thus the presence of particular kinds of domains into the protein structure could be of capital importance for the prediction of its function. In order to account for this source of information we included data published in [16]. This dataset has been processed in order to provide two types of information: the presence/absence of a particular protein domain in the proteins encoded by genes comprised in the dataset and the E-value assigned to each gene product to a collection of profile-HMMs,

each of which trained on a specific domain family. The E-values have been obtained by the HMMR software toolkit (`http://hmmer.janelia.org`).

The activation of a gene (and its functional products) is strictly regulated in cell in order to avoid interference between molecular processes, and this regulation is in part realized by modulating the transcriptional state of the gene. Genes involved in the realization of the same biological process are expected to show some similarities in their expression profiles. We thus included into our experiment a dataset obtained by the integration of microarray hybridization experiments published in [17] [18]. The main data sets used in the experiments are summarized in Tab. 1.

Table 1. Datasets

Data code	Dataset	examples	features
L_1	Protein domain binary	3529	4950
L_2	Protein domain log-E	3529	5724
L_3	Gene expression	4532	250
L_4	PPI - BioGRID	4531	5367
L_5	PPI - vonMering	2338	2559
L_6	Pairwise similarity	3527	6349

3.2 The Functional Catalogue (*FunCat*)

In order to associate each of the genes constituting the aforementioned datasets, we used functional annotations collected in the Functional Catalogue (FunCat) database [8], version (2.1), initially developed at MIPS during the early stages of sequencing of the yeast genome. The Functional Catalogue is constituted by hierarchically structured controlled vocabulary of functional categories. FunCat is the natural choice for our experiments, since it was originally developed to describe yeast functional processes.

In order to reduce the number of classification tasks required by the experimental setting we choose to consider only the first level FunCat classes. In other words, we selected the roots of the trees of the FunCat forest (that is the most general and wide functional classes of the overall taxonomy). We also removed by the list of the target functional classes all the classes represented by less than 20 genes. This corresponds to restrict our classifications to only 16 functional classes:
01:METABOLISM
02:ENERGY
10:CELL CYCLE AND DNA PROCESSING
11:TRANSCRIPTION
12:PROTEIN SYNTHESIS
14:PROTEIN FATE (folding,modification,destination)

16:PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)
18:REGULATION OF METABOLISM AND PROTEIN FUNCTION
20:CELLULAR TRANSPORT AND TRANSPORT ROUTES
30:CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM
32:CELL RESCUE, DEFENSE AND VIRULENCE
34:INTERACTION WITH THE ENVIRONMENT
40:CELL FATE
41:DEVELOPMENT(Systemic)
42:BIOGENESYS OF CELLULAR COMPONENTS
43:CELL TYPE DIFFERENTIATION

3.3 Base learners tuning and generation of optimized classifiers

The construction of an ensemble of classifiers trained to perform functional classification using heterogeneous data (as for any ensemble of classifiers) requires the definition of two key points: the way the base learners have to be trained and a strategy to combine the output of the different learner components. This section is dedicated to the former question while the second one has just be treated in Sect. 2.

Being the main objective of this experiment the evaluation of different strategies of fusion of the continuous output produced by different classifiers on heterogeneous datasets, we chose the simplest way to adapt the single sources of data: the intersection between all the datasets. This led to the definition of a common set constituted by 1901 genes.

For each of the 16 target functional classes we tuned the base learners as binary classifiers (thus labeling samples as belonging to the "target functional class" or to "other functional classes") using a classical inner cross-validation tuning scheme. More precisely, each dataset was split into a training set and a test set (composed, respectively, by the 70% and 30% of the available samples) and the resulting training set was furtherly split into 3 balanced folds, meaning that the proportion of positive and negative samples constituting each fold was kept equal for each fold.

The balanced folds have been used to perform a 3-folds cross validation for model selection. The averaged accuracy, precision, recall and F-measure across folds were collected for each combination of a list of tuning parameters. We chose RBF gaussian kernels for all the training tasks involved in the experiment, tuning each SVM for a cost ranging from 10^{-2} to 10^{2} and a value of sigma varying in the same range. During the tuning stage we experienced problems in tuning the learners dedicated to the classification of the Pairwise similarity dataset, for which only negative classifications were produced in 11 out of 16 learning tasks. We thus changed the tuning setting for this dataset by using a polynomial kernel and varying the degree hyperparameter from 2 to 5 while keeping the cost varying from 10^{-2} to 10^{2}.

Among the commonly used performance metrics suitable to drive the optimization process, considering that negative examples are largely less than positives, we decided to tune the base learners by choosing the set of parameters producing the maximum averaged F-measure during the tuning stage. Once defined the best set of parameters associated to each learner in each learning task, we used them to train an optimal model on the whole training set.

The generalization performances have been estimated on the separated test set.

For the experiments we used the *Lagrange* cluster composed by 208 nodes equipped with Intel Xeon 3.16 GHz QuadCore processors and 16 GB of RAM memory at each node (`http://www.cilea.it`).

4 Results

The performances obtained in the learning tasks associated to the prediction of the FunCat functional classes are reported in Tab. 2.

The table reports the F-measures obtained through the evaluation of the test set (570 genes) using the best models selected by internal 3-folds cross-validation. The first column refers to the FunCat identifiers of first-level functional classes. The next 6 columns (from L_1 to L_6) correspond to single SVMs trained respectively on the six datasets described in Tab. 1. L_{avg} represents the averaged results of the single SVMs across the six datasets, and the three last columns refers respectively to the weighted linear, logarithmic and decision template ensembles. The performances of the best performing single learner and the best performing ensemble are highlighted in boldface.

Note that among the first level FunCat functional classes, the "DEVELOPMENT(Systemic)" class, (class ID: 41) is not reported because, after the intersection of the data sources described in Tab. 1, it fails to reach the minimum amount of positive samples (20 genes belonging to the target functional class) required by our experimental protocol.

Looking at the last row of Tab. 2, we see that, on the average, data integration methods through ensembles provide better results than single SVMs trained on homogeneous bio-molecular data, independently of the applied combination rule. In particular decision templates largely outperform both the single SVMs and the other ensembles. Performances of weighted and logarithmic ensembles are quite comparable, better than the average single SVM and in most cases better than the single SVM trained on a single data set.

F-measure performances are summarized in Fig. 1: all ensemble methods outperform the F-measure obtained, on the average, by the single SVMs. The best single SVM for each task outperforms weighted linear and logarithmic ensembles, but decision templates are in most cases better than the best single SVM.

Table 2. F-measures computed on the test sets (see text for more details)

FunCat class	L_1	L_2	L_3	L_4	L_5	L_6	L_{avg}	E_{lin}	E_{log}	E_{DT}
01	0.6240	0.6486	0.4854	0.6461	0.5283	**0.7576**	0.6150	0.7835	**0.7860**	0.7845
02	0.2258	0.3478	0.2941	0.2318	0.3125	**0.4000**	0.3020	0.2857	0.3125	**0.4324**
10	0.5240	**0.6819**	0.1916	0.4059	0.3800	0.5963	0.4632	0.5887	0.5887	**0.6666**
11	0.5607	**0.7213**	0.2395	0.4397	0.4524	0.5693	0.4971	0.5673	0.5673	**0.6722**
12	0.6060	0.6616	0.3207	0.4793	**0.7361**	0.5181	0.5536	**0.6814**	0.6412	0.6715
14	0.4331	0.5622	0.4221	**0.6234**	0.2772	0.6191	0.4895	0.6776	0.6581	**0.6846**
16	0.3771	**0.4661**	0.2561	0.4086	0.2040	0.5146	0.3710	0.5217	0.4978	**0.5543**
18	0.0000	0.0526	0.1764	0.2352	0.0000	**0.2857**	0.1249	0.2424	0.2424	**0.3333**
20	0.4457	**0.6461**	0.2733	0.4588	0.1492	0.4802	0.4088	**0.5828**	0.5212	0.5465
30	0.0975	0.3913	0.1818	0.2702	0.0000	**0.4266**	0.2279	0.2285	0.2352	**0.5769**
32	0.2278	0.2650	0.2025	**0.3146**	0.0000	0.2684	0.2130	0.1842	0.1351	**0.2500**
34	0.3023	**0.3544**	0.0909	0.2133	0.0000	0.1834	0.3023	0.1764	0.1764	**0.4509**
40	0.2307	0.2745	0.1250	0.1250	0.0000	**0.3000**	0.1758	0.1304	0.1304	**0.3409**
42	0.4129	**0.5524**	0.0847	0.0344	0.1068	0.4052	0.2660	0.4736	0.3333	**0.5279**
43	0.4150	**0.6016**	0.2000	0.2000	0.0000	0.4153	0.3053	0.3956	0.3414	**0.4600**
AVG	0.3655	**0.4818**	0.2363	0.3391	0.2098	0.4493	0.3544	0.4347	0.4111	**0.5302**

5 Discussion

The prediction of the functional class of genes using heterogeneous data sources is among the most difficult problems in bioinformatics. The difficulty of the task comes mainly from the unavoidable not so strict definition of the "biological process" entity, due to the high number of interconnections linking biological processes, and to the different relevance of diverse biomolecular datasets with respect to different functional classes. A dataset providing critical information for the prediction of a particular functional class could merely represent noise in a classification task targeting a different functional class.

As shown by data reported in tab. 2, the best performing ensemble system outperforms the average performances obtained by the base learners in all the functional classification tasks. The winner combination method in 13 out of 15 test is the ensemble based on decision templates. If compared with results obtained by the best performing base learner, the best performing ensemble systems win on 8 out of 15 test. This comes as a certain surprise because of the presence in these ensembles of base learners with very poor performances (as reported in the cases of tasks aimed to predict the 20 (CELLULAR TRANSPORT AND TRANSPORT ROUTES), 32 (CELL RESCUE DEFENSE AND VIRULENCE) and 42 (BIOGENESIS OF CELLULAR COMPONENTS) FunCat classes. In this work we did not explore the effects on ensembles performance of methods that selects subsets of base learners. Indeed, in this preliminary test, we were mainly interested in the eval-

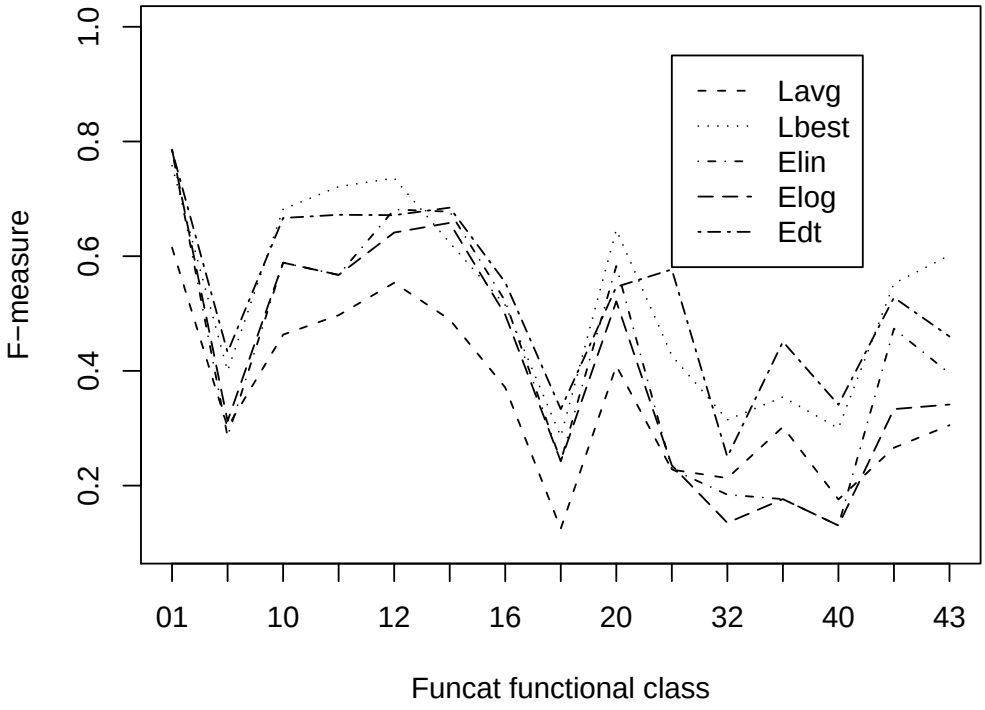

Fig. 1. Comparison of the F-measures achieved in gene prediction: Lavg stands for the average across SVM base learners; Lbest for the best single SVM; Elin, Elog, Edt for weighted linear, logarithmic and decision template ensembles.

uation of the potential benefits introduced by the use of data fusion methods in complex functional prediction tasks, but base learner selection approaches will be the object of future investigations.

Despite the expected negative effects of poor performing base learners on the ensemble systems performances, the ensemble systems are surprisingly robust as demonstrated by the performances obtained in the test prediction of the class 34 (INTERACTION WITH THE ENVIRONMENT). In this classification task, despite the presence of 4 out of 6 base learners with performance around 20% or less, and the remaining learners with a F-measure of 0.3023 and 0.3544 (L_1 and L_2, respectively), the decision templates based ensemble system outperform the best base learner of about 10%. This could be inter-

preted as the ability of the different data sources to provide diverse pieces of information.

The protein involved in the interaction with the environment are characterized by very peculiar chemical and physical properties, as in the case of cell membrane spanning proteins (expected to be over represented in this functional class), which are composed by hydrophilic and hydrophobic alternate regions, making them easily detectable by protein-domain detection methods and sharing features making them the objective of evolutionary pressures easily detectable by local alignment methods. The ability of ensemble methods to correctly predict this FunCat functional class thus comes as a little surprise, being the aforementioned source of information well represented by 3 out of 6 data sources (the Protein domain binary, Protein domain Log-E and Pairwise similarity, Tab. 1).

It should be also noted that the realization of the complex pathways enabling the cell to correctly interact with its environment requires the interaction between a high number of proteins, making the proteins interaction data sources potentially informative. This indicates that some poor performances (either of the base learners or the ensemble systems) could be explained by the absence of datasets containing relevant information with respect to the specific functional prediction task. We thus plan to extend the number of the datasets to be included in further analyses.

6 Conclusions

In this work we investigated the impact on yeast genes functional classification performances of data fusion methods based on ensemble methods. Our experiments consisted in the integration (by mean of a simple intersection procedure) of 6 different data sources and in the training (using standard tuning protocols) of 6 SVMs. We then tested linear weighted average, logarithmic weighted average and the decision templates techniques to combine the output of the 6 base learners and we evaluated the performance of the single learners and of the ensemble systems. The aforementioned protocol was repeated (in separated binary classification tasks) for each FunCat functional class performed on the common set of genes shared by all the data sources to be integrated.

Our experiments demonstrated the potential benefits introduced by the use of ensemble-based prediction systems in functional classification of genes. The ensemble systems were able to outperform the averaged performances of base learners in all the gene function prediction tasks, and were able to outperform the best performing base learner in 8 out of 15 classification tasks. Considering that poor performances of base learners are functional class specific, by appropriately combining subsets of base learners for each specific FunCat class we could in principle improve the performances of SVM ensembles. In conclusion, the results obtained with simple combination strategies

show that heterogeneous data integration through ensemble methods represents a valuable research line in gene function prediction.

Acknowledgments

The authors gratefully acknowledge partial support by the PASCAL2 Network of Excellence under EC grant no. 216886. This publication only reflects the authors' views.

References

1. Pavlidis, P., Weston, J., Cai, J., Noble, W.: Learning gene functional classification from multiple data. J. Comput. Biol. **9** (2002) 401–411
2. Troyanskaya, O., et al.: A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *saccharomices cerevisiae*). Proc. Natl Acad. Sci. USA **100** (2003) 8348–8353
3. Lanckriet, G., De Bie, T., Cristianini, N., Jordan, M., Noble, W.: A statistical framework for genomic data fusion. Bioinformatics **20** (2004) 2626–2635
4. Roli, F., Kittler, J., Windeatt, T.: Multiple Classifier Systems, Fifth International Workshop, MCS2004. Volume 3077 of Lecture Notes in Computer Science. Springer-Verlag, Berlin (2002)
5. Guan, Y., Myers, C., Hess, D., Barutcuoglu, Z., Caudy, A., Troyanskaya, O.: Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biology **9**(S2) (2008)
6. Vapnik, V.N.: The nature of Statistical Learning Theory. Springer, New York (1995)
7. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Smola, A., Bartlett, P., Scholkopf, B., Schuurmans, D., eds.: Advances in Large Margin Classifiers. MIT Press, Cambridge, MA (1999)
8. Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M., Mewes, H.: The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research **32**(18) (2004) 5539–5545
9. Kuncheva, L., Whitaker, C.: Measures of diversity in classifier ensembles. Machine Learning **51** (2003) 181–207
10. Kuncheva, L.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, New York (2004)
11. Kuncheva, L., Bezdek, J., Duin, R.: Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition **34**(2) (2001) 299–314
12. Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. Nucleic Acids Res. **34** (2006) D535–D539
13. vonMering, C., et al.: STRING: a database of predicted functional associations between proteins. Nucleic Acids Research **31** (2003) 258–261

14. Smith, T., Waterman, M.: Identification of common molecular subsequences. Journal of Molecular Biology **147** (1981)
15. Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D.: Basic local alignment search tool. Journal of Molecular Biology **215** (1990)
16. Finn, R., Tate, J., Mistry, J., Coggill, P., Sammut, J., Hotz, H., Ceric, G., Forslund, K., Eddy, S., Sonnhammer, E., Bateman, A.: The Pfam protein families database. Nucleic Acids Research **36** (2008) D281–D288
17. Gasch, P., et al.: Genomic expression programs in the response of yeast cells to environmental changes. Mol.Biol.Cell **11** (2000) 4241–4257
18. Spellman, P., et al.: Comprehensive identification of cell cycle-regulated genes of the yeast saccharomices cerevisiae by microarray hybridization. Mol. Biol. Cell **9** (1998) 3273–3297