Soft Comput
DOI 10.1007/s00500-010-0599-6

FOCUS

Comparing early and late data fusion methods for gene expression prediction

Matteo Re

Abstract The most basic molecular mechanism enabling a living cell to dynamically adapt to variation occurring in its intra and extracellular environment is constituted by its ability to regulate the expression of many of its genes. At biomolecular level, this ability is mainly due to interactions occurring between regulatory motifs located in the core promoter regions and the transcription factors. A crucial question investigated by recently published works is if, and at what extent, the transcription patterns of large sets of genes can be predicted using only information encoded in the promoter regions. Even if encouraging results were obtained in gene expression patterns prediction experiments the assumption that all the signals required for the regulation of gene expression are contained in the gene promoter regions is an oversimplification as pointed out by recent findings demonstrating the existence of many regulatory levels involved in the fine modulation of gene transcription levels. In this contribution, we investigate the potential improvement in gene expression prediction performances achievable by using early and late data integration methods in order to provide a complete overview of the capabilities of data fusion approaches in a problem that can be annoverated among the most difficult in modern bioinformatics.

Keywords Weighted averaging · Decision templates · Vector space integration · Early fusion · Late fusion · Decision fusion · Data integration · Gene expression prediction

M. Re (✉)
Dipartimento di Scienze dell'Informazione, DSI, Universitá degli studi di Milano, via Comelico 39, Milan, Italy
e-mail: re@dsi.unimi.it

Published online: 21 March 2010

1 Introduction

The information required for the construction of proteins, the main players involved into the realization of the complex set of biochemical and metabolic reactions occurring in living cells, are encoded in the DNA in form of informational units called genes. The regulation of gene expression is of capital importance in order to ensure the presence of the required proteins at the right moment and in specific subcellular locations.

A great part of the ability to regulate gene expression at cellular level is due to the presence of many signals encoded in the core promoter, a relatively small region located immediately upstream the transcription start site (TSS). According to classical biomolecular models, gene expression is regulated by proteins known as transcription factors (TFs) that interact with cis-regulatory elements, the transcription factor binding sites (TFBS), located in the promoter regions. Only in response to a specific set of environmental conditions, the right combination of TFs bind the TFBSs, and this event enables the cellular transcriptional machinery to start the transcription of the gene.

The complexity of these gene expression regulation models relies on the combinatorial nature of the TF action, since the binding of a specific TF to a specific core-promoter sequence can both enhance or silence the transcription of the regulated gene, according to complex regulatory networks that are, at today, only partially understood.

Recently published works demonstrated that connectivity maps based on gene-expression signatures, data collected in presence/absence of perturbagens and clinical information are tool of election for the characterization of functional associations among diseases, gene perturbations

and drug action Lamb et al. (2006). Other recently proposed methods are able to predict Drug mode of action (MOA) of novel compounds using gene expression profiles (GEPs) Iorio et al. (2009) and to explain genome wide expression data by focusing on gene sets, that is, groups of genes that share common biological function, chromosomal location, or regulation Subramanian et al. (2005). While these methods are of crucial importance in clinical and pharmaceutical investigations, being them able to detect functional associations between genes,chemicals and diseases using expression profiles, they are not suitable for the prediction of co-expressed groups of genes without the direct evaluation of expression data.

A key point required for the elucidation of the transcriptional regulation mechanisms is the definition of the minimal set of DNA regulatory signatures (comprising combinations of TFBSs, evolutionary constraints, epigenetic modifications, and many others) responsible for specific expression patterns characterizing co-regulated genes.

In a recently published work Beer and Tavazoie (2004) the authors tried to predict the expression class of yeast genes using only information encoded in the promoter regions. The expression classes were obtained using a clustering algorithm Hartigan (1975) to find genes that are co-expressed across a broad range of conditions. The underlying assumption in this work is that genes sharing similar expression patterns have to share also a common set of signals in their promoter regions. The authors were able to predict the expression pattern of the genes in many and stress conditions, using only the signals encoded in the promoter regions (the TFBSs, their location and orientation) achieving a 73% accuracy.

Despite the short cis-regulatory sequences contained in the promoter regions are key players in the regulation of the first step of gene transcription, other mechanisms are involved the regulation of gene expression. Recently, Millar and Grunstein (2006) demonstrated that post translational histone modification are able to modulate the expression pattern of genes. Other useful information can be obtained by investigating the conservation of the TFBSs across different but phylogenetically related organisms McIsaac et al. (2006). This approach is motivated by the observation that the strength of the selective pressures acting during evolution on cis-regulatory motifs sited in the core promoters could help to filter out noisy and a specific TFBS.

Histone modifications and phylogenetic conservation are only two of the potentially useful source of information for gene expression prediction as recent advances in biotechnologies resulted in the last years into an ever increasing number of biomolecular datasets available in the public domain.

In order to effectively exploit these information for gene expression prediction a key problem is the integration of heterogeneous biomolecular data. Data fusion approaches can be roughly classified according to the moment in which the integration of heterogeneous data occurs. In early integration methods the integration is performed at feature level, as in the case of the direct "vector-space integration" (VSI) in which different vectorial data are concatenated desJardins et al. (1997) and then used to train a final classifier. Kernel methods, by exploiting the closure property with respect to the sum, represents another valuable research direction for the integration of biomolecular data Lanckriet et al. (2004).

All these methods suffer of limitations and drawbacks, due to their limited scalability to multiple data sources [as in the case of Kernel integration methods based on semidefinite programming Lanckriet et al. (2004)], to their limited modularity when new data sources sources are added (e.g. vector-space integration methods), or when data are available with different data type representations (e.g. functional linkage networks and vector-space integration). A possible alternative approach is based on ensemble methods.

In late fusion methods, as in the case of ensemble systems, a single learner is trained for any available datasource and the base learners outputs are then converted into a common form resulting into an intermediate feature space in which a suitable rule can be applied to make a final decision. To our knowledge, this is the first work devoted to the investigation of performances achievable in gene expression prediction by using early and late data integration methods. In this contribution, we compare the effectiveness of an early fusion method (direct vector space integration) and several late integration approaches: the classical weighted integration (using two different weighting schemes) and the Decision Templates combiner Kuncheva et al. (2001) in order to provide an overview of capabilities of multiple classifier systems in the integration of heterogeneous biomolecular data sources for the prediction of gene expression.

2 Heterogeneous data integration: the early and the late fusion approaches

2.1 Early fusion by vector space integration

The simplest form of heterogeneous data integration is to concatenate the features collected for each gene in all the available datasets in a fixed-length vector and then feed the resulting collection of vectors into a classification algorithm Pavlidis et al. (2002). The vector-space integration (VSI) is suitable for data integration independently from

the structure of the involved dataset and has the advantage of simplicity. VSI is suffering of biases due to the different length of the concatenated vectors and is not able to incorporate much domain knowledge being each type of data treated identically Noble and Ben-Hur (2007). In our experiments, we normalized the data with respect to the mean and standard deviation, separately for each data set.

2.2 Reasons motivating the use of the late integration approach based on ensemble systems

There are several reasons to apply ensemble methods in the specific context of genomic data fusion for gene expression prediction. At first, continuous advances in high-throughput biotechnologies provide new types of data, as well as updates of existing biomolecular data available for gene expression prediction.In this context, ensemble methods are well-suited to embed new types of data or to update existing ones by training only the base learners devoted to the newly added or updated data, without retraining the entire ensemble. Moreover most ensemble methods scale well with the number of the available data sources, and problems that characterize other data fusion approaches are thus avoided. Using vectorial data for different sources there is no bias in the integration of large and small or sparse and dense vectors. More in general diverse types of data (e.g. sequences, vectors, graphs) can be easily integrated, because with ensemble methods the integration is performed at decision level. Data fusion of heterogeneous biomolecular data sources can be effectively realized by means of ensemble systems composed by base learners trained on different datasets, and then combining their outputs to compute the consensus decision.

2.3 The simplest form of late fusion integration: the weighted average

In the context of gene expression classification, we need to estimate of the reliability of the prediction. To this end, we use SVMs, with probabilistic output obtained by applying a sigmoid fitting to their output Lin et al. (2007). Thus a trained base 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 n base learners, each trained on a different type of biomolecular data, using a suitable combining function g to compute the overall probability μ_j for a given class ω_j:

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

A simple way to integrate different biomolecular data sources is represented by the weighted linear combination rule:

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

The weights are usually computed using an estimate of the overall accuracy of the base learners, but for gene function prediction, where the functional classes are largely unbalanced (positive examples are largely less than negative ones), we choose the F-measure (the harmonic mean between precision and recall). We consider two different ways to compute the weights:

$$w_t^{\mathrm{I}} = \frac{F_t}{\sum_{t=1}^{n} F_t} \qquad w_t^{\log} \propto \log \frac{F_t}{1 - F_t} \tag{3}$$

The w_t^{I} weights are obtained by a linear combination of the F-measures, and $w_t^{\log}$ by a logarithmic transformation. Independently of the choice of the weights the decision $D_j(\mathbf{x})$ of the ensemble about the class ω_j is taken using the estimated probability μ_j (Eq. 2):

$$D_j(\mathbf{x}) = \begin{cases} 1, & \text{if } \mu_j(\mathbf{x}) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where output 1 correspond to positive predictions for ω_j and 0 to negatives.

2.4 Late integration accounting for systematic errors in base learners outputs: the Decision Templates combiner

Certain types of biomolecular data can be informative for some expression classes, but uninformative for others. Hence it would be helpful to take into account whether certain types can be informative or not, depending on the class to be classified. To this end *Decision Templates* Kuncheva et al. (2001) can represent a valuable approach. The main idea behind decision templates consists in comparing a "prototypical answer" of the ensemble for the examples of a given class (the template), to the current answer of the ensemble to a specific example whose class needs to be predicted (the decision profile).

More precisely, the decision profile $\mathrm{DP}(\mathbf{x})$ for an instance $\mathbf{x}$ is a matrix composed by the $d_{t,j} \in [0, 1]$ elements representing the support given by the tth classifier to class ω_j. 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 :

$$\mathrm{DT}_j = \frac{1}{|\mathbf{X}_j|} \sum_{\mathbf{x} \in \mathbf{X}_j} \mathrm{DP}(\mathbf{x}) \tag{5}$$

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

Table 1 Datasets

Code	Dataset	Examples	Features	Description
D_{tavR}	Beer motif scores real	2,587	666	Beer motif scores (Real) from Beer and Tavazoie (2004)
D_{tavB}	Beer motif scores binary	2,587	666	Beer motif scores (binary) from Beer and Tavazoie (2004)
$D_{histmod}$	Histone modification scores	2,580	22	Histone modification scores collected from the ChromatinDB O'Connor and Wryck (2007) database
D_{phylo}	Motifs conservation scores	2,492	121	Motifs conservation scores produced using the PhyloCon algorithm McIsaac et al. (2006)

$$\mathcal{S}_j(\mathbf{x}) = 1 - \frac{1}{n \times c} \sum_{t=1}^{n} \sum_{k=1}^{c} [\mathrm{DT}_j(t,k) - d_{t,k}(\mathbf{x})]^2 \quad (6)$$

The final decision of the ensemble is taken by assigning a test instance to a class with the largest similarity:

$$D(\mathbf{x}) = \arg\max_j \mathcal{S}_j(\mathbf{x}) \quad (7)$$

In our experimental setting we consider dichotomic problems, because a gene may belong or not to a given expression class, thus obtaining two-columns decision template matrices. It is easy to see that with dichotomic problems the similarity $(\mathcal{S}_1)$ (Eq. 6) for the positive class and the similarity $(\mathcal{S}_2)$ for the negative class become:

$$\mathcal{S}_1(\mathbf{x}) = 1 - \frac{1}{n} \sum_{t=1}^{n} [\mathrm{DT}_1(t,1) - d_{t,1}(\mathbf{x})]^2 \quad (8)$$

$$\mathcal{S}_2(\mathbf{x}) = 1 - \frac{1}{n} \sum_{t=1}^{n} [\mathrm{DT}_2(t,1) - d_{t,1}(\mathbf{x})]^2 \quad (9)$$

where DT_1 is the decision template for the positive class and DT_2 for the negative one. The final decision of the ensemble for a given functional class is:

$$D(\mathbf{x}) = \arg\max_{\{1,2\}} (\mathcal{S}_1(\mathbf{x}), \mathcal{S}_2(\mathbf{x})) \quad (10)$$

3 Experimental setup

We choose to perform our experiments using the data provided in Beer and Tavazoie (2004) supplemental materials and two additional datasets. In Beer and Tavazoie (2004), the authors used not only the matching scores of the motifs in the promoter regions but also their location and orientation. In this experiment, we used only the matching scores. The motifs scores used as indicators of the presence/absence of the TFBSs in the gene promoters in Beer and Tavazoie (2004) were used in the form provided by the authors and in form of binary indicators.

We also included two additional datasets collected, respectively, from the ChromatinDB database O'Connor and Wryck (2007) and from McIsaac et al. (2006) supplemental material.

Genome-wide Chromatin Immuno Precipitation (ChIP) data for 22 different histone modifications were downloaded from ChromatinDB O'Connor and Wryck (2007). We extracted from ChromatinDB all the available data inherent to ChIP data annotated in the genomic regions corresponding to all the annotate *Saccharomyces cerevisiae* gene promoters. Thanks to the data preprocessing policies adopted in the development of the ChromatinDB database the data, collected from literature, are available both in raw and normalized form in order to avoid biases introduced by the differences in the experimental setup under which the data were originally produced. All the data collected from ChromatinDB were retrieved in normalized form.

The last dataset involved in our experiments is based on the conservation scores produced by the PhyloCon algorithm McIsaac et al. (2006). The authors provided these data in form of three tables of motifs scores expressing the conservation level of the motifs annotated in *S. cerevisiae* promoters produced by comparative genomics methods based on the comparison of orthologous promoters pairs. The data are provided in form of three table dedicated to low, moderately and highly conserved motifs. The PhyloCon data were merged into an unique table expressing the conservation level of all the TFBSs in form of discrete and ordered indicators ranging from 0 (not conserved) to 3 (highly conserved).

The expression data used in Beer and Tavazoie (2004) for the clustering analysis resulting in the definition of the 49 expression classes and constituting the labels in our experiments are published in Gasch et al. (2000), Spellman et al. (1998). The main characteristics of the data sets used in the experiments are summarized in Table 1.

We considered yeast genes common to all data sets (2490), and we associated them to the expression classes reported in Beer and Tavazoie (2004). The investigated classification problems are affected by a severe unbalance between positives and negatives examples ranging the number of positive examples from 5.0 to 0.5% of the available data depending on the considered expression

Table 2 Number of positives and negatives examples in the 41 investigated expression classes

ExprClass	Posi	Nega	TRpos	TRneg	TEpos	TEneg
01	124	2,366	87	1,656	37	710
02	108	2,382	76	1,667	32	715
03	104	2,386	73	1,670	31	716
04	104	2,386	73	1,670	31	716
05	81	2,409	57	1,686	24	723
37	31	2,459	22	1,721	9	738
38	29	2,461	20	1,723	9	738
39	32	2,458	22	1,721	10	737
40	27	2,463	19	1,724	8	739
41	27	2,463	19	1,724	8	739

Each row of the table lists the code of the expression classes (ExprClass) corresponding to the labels defined in Beer and Tavazoie (2004), the number of positive and negatives points for each class (Posi and Nega columns), and the number of positive and negative points in the training (TR) and test (TE) sets (last four columns). The rows are sorted according to the number of positive examples. From the entire set of the 41 considered expression classes, the table lists only the top 5 and bottom 5

Table 3 Unbalanced setup

Metric	E_{lin}	E_{log}	E_{dt}	*VSI*	D_{avg}	D_{tav}
F	0.1087	0.1419	0.2683	0.1230	0.1077	0.2094
acc	0.9233	0.9481	0.9119	0.9769	0.9768	0.9773
prec	0.2440	0.3332	0.4361	0.2949	0.2160	0.4293
rec	0.1409	0.1372	0.2563	0.0874	0.0802	0.1559

Late fusion methods, VSI, average performances of base learners and performances of D_{tav}: average *F*-measure, accuracy, precision and recall evaluated by multiple hold-out

Table 4 Balanced setup

Metric	E_{lin}	E_{log}	E_{dt}	*VSI*	D_{avg}	D_{tavR}
F	0.7891	0.7891	0.7913	0.6844	0.6989	0.7832
acc	0.7914	0.7915	0.7931	0.6806	0.6604	0.7850
prec	0.7993	0.7992	0.7999	0.6857	0.6490	0.7896
rec	0.7903	0.7904	0.7904	0.6967	0.8128	0.7910

Ensembles of learning machines, average performances of base learners and performances of D_{tav} : average *F*-measure, accuracy, precision and recall computed by multiple hold-out techniques

class (see Table 2). In order to avoid classification tasks with a too low number of positive examples the 8 smallest expression classes were excluded from our experiments resulting in 41 binary classification problems.

The learning problem was split in 41 binary classification tasks in which each gene was predicted as belonging or not to the considered expression class.

Each dataset was split into a training set and a test set (composed, respectively, by 70 and 30% of the available samples). We performed a threefold stratified cross-validation on the training data for model selection: we computed the *F*-measure across folds, while varying the parameters of gaussian kernels (both σ and the C regularization term, ranging from 10^{-5} to 10^{5}).

Classification performances of the component classifiers, the ensemble systems and VSI have been evaluated using a multiple hold-out scheme based on five replicates of the aforementioned training and testing procedure. The collected test sets classification performances have been averaged across all the replicates.

In order to evaluate the gain in prediction performances achievable by data integration methods in presence and in absence of the problems due to the unbalance between positives and negatives examples we repeated the entire procedure using artificially balanced datasets constituted by all the positive examples belonging to the considered expression class and the same amount of negative examples randomly chosen from the remaining expression classes. We adopted many performances evaluators, instead of the Accuracy used by Beer and Tavazoie (2004). Our choice is motivated by the large unbalance between positive and negative examples that characterizes the investigated prediction problems: indeed on the average only a small subset of the available genes is annotated to each expression class (see Table 2). We compared the performances of single gaussian SVMs trained on each data set with those obtained with the late fusion approaches described in Sects. 2.3 and 2.4 and with the performances obtained with the early fusion method, VSI, described in Sect. 2.1. We normalized the data with respect to the mean and standard deviation, separately for each data set.

4 Results

Results obtained in the unbalanced learning tasks are summarized in Table 3. The table shows the average *F*-measure, accuracy, precision and recall across the 41 selected gene expression classes, obtained through the evaluation of the test sets (each constituted by 747 genes). The performances are estimated using a multiple hold-out based on five replicates and the final test sets performances are averaged. The three first columns are dedicated to the late fusion methods and refer, respectively, to the weighted linear, weighted logarithmic and decision template ensembles (see Sects. 2.3 and 2.4), VSI represent the averaged results obtained by the early integration method Vector Space integration, D_{avg} represents the averaged results of the single SVMs across the four datasets, and D_{tavR} represents the single SVM trained using data provided by Tavazoie and colleagues (Table 1). Table 4 shows the same results obtained in the balanced learning tasks.

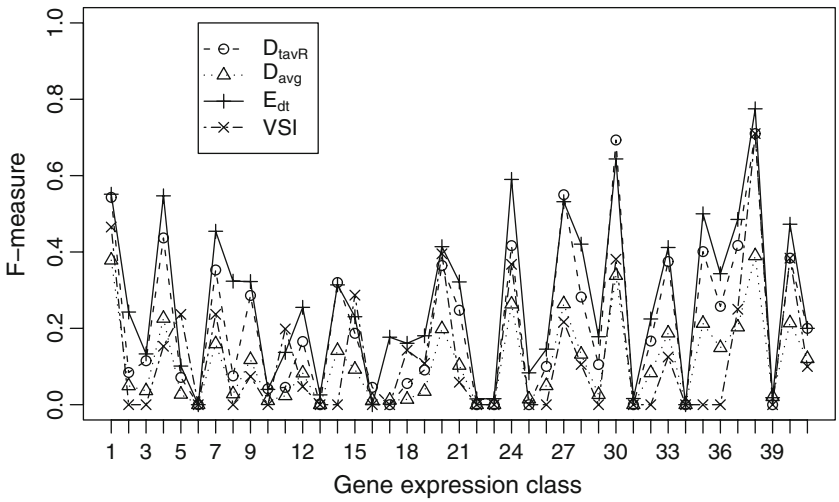

Fig. 1 Unbalanced setup: comparison of the averaged F-measures achieved in gene expression prediction. D_{avg} stands for the average across the base learners, D_{tavR} for the best component classifier, *VSI* for the early integration method and E_{dt} for the best performing late integration method

Under the unbalanced experimental setup (see Table 3), the large accuracies are due to the concurrent failure of the component classifiers in the learning problems (meaning that in many classification tasks all the test instances were predicted as negatives), and the large unbalance in the data. According to the collected F-measures we observe that data integration methods are able on the average to outperform the component classifiers independently by the considered data fusion approach. Interestingly this trend is not confirmed in the comparison of the data fusion methods with results obtained by the best performing component classifier: D_{tavR}. In particular 2 out of 3 late fusion methods (E_{lin} and E_{log}) are on the average unable to outperform D_{tavR}.

The early fusion method, VSI, was also unable to outperform D_{tavR} but is able, according to the observed F-measure, to outperform the late fusion approach based on the weighted averaging using linear weights (E_{lin}). The only data fusion method able to outperform the best component classifier (in 33 over 41 classification tasks) is the Decision Templates combination rule. The ability of E_{dt} to outperform D_{tavR} is also confirmed looking at the Precision and the Recall.

In this extremely difficult classification test, the collected results confirmed the ability of the Decision Templates combiner to learn not only from correct predictions but also from the wrong ones exploiting the different patterns in the errors produced during the classification of the positive and negative instances.

According to the collected F-measures averaged for each gene expression class across the performed replicates, the data fusion methods $E_{\text{lin}}, E_{\text{log}}, E_{dt}$ and *VSI* were able to outperform the best component classifier (D_{tavR}), respectively, 4, 2, 33, and 6 times under the unbalanced setup, indicating that in critically difficult gene expression prediction problems the Decision Templates ensemble system is the safer choice.

In order to evaluate the impact on classification performances of the severe unbalance affecting the data we repeated the entire experiment by random sampling, in each classification task, a number of negative instances equal to the number of the positive ones. The summary of the averaged results collected in the artificially balanced gene expression prediction tasks are reported in Table 4. The table shows the average F-measure,accuracy, precision and recall across the 41 selected gene expression classes, obtained through the evaluation of the test sets. The table has the same structure of Table 3.

Looking at the values presented in Table 4 and considering the F-measure, we see that, on the average, in the artificially balanced setup late fusion methods realized by using ensemble methods provide better results than single SVMs, independently of the applied combination rule. In particular Decision Templates achieved the best average F-measure albeit the performances are quite similar for all the tested combination methods. The early fusion method (VSI) was unable to outperform the averaged performances of the component classifiers.

Considering the averaged accuracies the data fusion methods are still able to outperform all the component SVMs. The observed trend is confirmed for Precision but not for the Recall: only the Decision Templates combiner was able to outperform all the component classifiers independently of the considered performance metric.

Under this balanced setup, and using the accuracy as performance metric, we outperformed the results obtained by Beer and colleagues (73% accuracy).

The averaged F-measures collected during the test sets evaluation under the unbalanced setup and the artificially balanced setup are reported in Figs. 1 and 2, respectively.

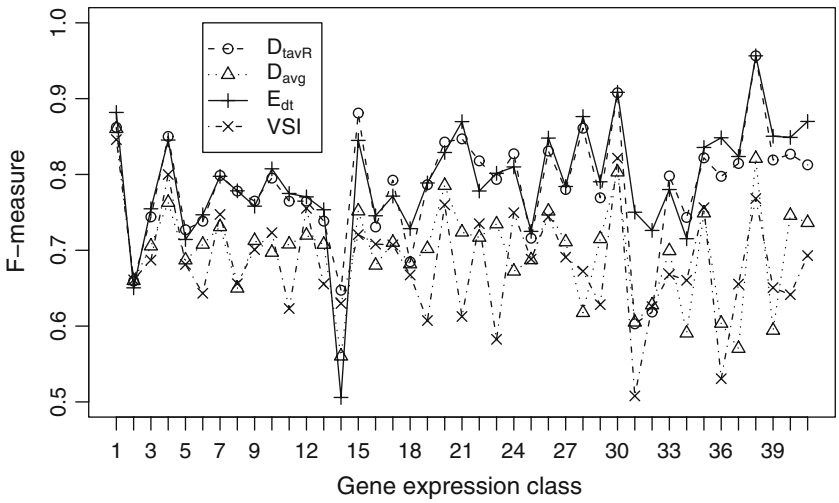

Fig. 2 Balanced setup: comparison of the averaged *F*-measures achieved in gene expression prediction. D_{avg} stands for the average across the base learners, D_{tavR} for the best component classifier, *VSI* for the early integration method and E_{dt} for the best performing late integration method

5 Conclusions

In this work, we compared the performances in yeast gene expression prediction of early and late data fusion methods. Many methods suitable for heterogeneous data integration (like Ada Boost and many others Friedman et al. 2000; Rosset et al. 2004; Zhu et al. 2004) are available but, in this preliminary investigation, we decided to systematically test simple integration methods in order to evaluate the potential of data fusion methods in gene expression prediction. A possible future direction could be the test of the performances of a more broad set of data integration methods in gene expression prediction problems.

Despite the extreme difficulty of the investigated classification problems due to a severe unbalance affecting the datasets involved in our experiments, our results clearly demonstrated the potential benefits in classification performances introduced by the use of relatively simple late integration methods.

Among the tested data integration approaches the VSI method obtained the worst classification performances indicating that early integration methods are not well suited for gene expression prediction problems. Our observations are supported both in the original severely unbalanced problem investigated by Tavazoie and colleagues and under an artificially balanced setup in which one of the investigated methods, the Decision Templates combiner, was also able to outperform the results presented in Beer and Tavazoie (2004) (73%) achieving a final 79% accuracy.

A reexamination of the original experiment performed by Beer and Tavazoie (2004) was recently published in Yuan et al. (2007). In agreement with our results Yuan and colleagues found that, using the same dataset published in Beer and Tavazoie (2004), better gene expression prediction performances can be achieved avoiding the use of the position of the regulatory motifs along the promoter regions and their orientation. Our results are also comparable with the results recently published in Pavesi and Valentini (2009) confirming that the proposed method is able to provide an overall classification accuracy that is comparable with other state of the art studies aimed to predict the expression class of co-regulated genes. Using performance measures well-suited to unbalanced problems, we also demonstrated that, in critically difficult gene expression prediction problems involving severely unbalanced datasets, the use of late integration methods and, in particular, the Decision Template combiner can improve the classification performances both in terms of precision and recall.

The results presented in this contribution, obtained with relatively simple combining methods, show the effectiveness of the proposed approach and demonstrates that data fusion realized using ensemble systems is a promising research line in gene expression prediction.

Acknowledgments The authors would like to gratefully acknowledge partial support by the PASCAL2 Network of Excellence under EC grant no. 216886. This publication only reflects the authors' views. The author would also like to expressly thank Giorgio Valentini for the examination of early versions of the manuscript.

References

Beer M, Tavazoie S (2004) Predicting gene expression from sequence. Cell 117

desJardins M et al (1997) Prediction of enzyme classification from protein sequence without the use of sequence similarity. In: Proceedings of the 5th international conference on intelligent systems for molecular biology. AAAI Press, Menlo Park, pp 92–99

Friedman J et al (2000) Additive logistic regression: a statistical view of boosting. Ann Stat 38(2):337–374

Gasch P et al (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11:4241–4257

Hartigan J (1975) Clustering algorithms. Wiley, New York
Iorio F et al (2009) Identifying network of drug mode action by gene expression profiling. J Comput Biol 16
Kuncheva LI et al (2001) Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognit 34(2):299–314
Lamb J et al (2006) The connectivity map: using gene-expression signatures to connect small molecules genes and diseases. Science 313
Lanckriet G et al (2004) A statistical framework for genomic data fusion. Bioinformatics 20:2626–2635
Lin H, Lin C, Weng R (2007) A note on Platt's probabilistic outputs for support vector machines. Mach Learn 68:267–276
McIsaac K et al (2006) An improved map of conserved regulatory sites map for *Saccharomyces cerevisiae*. BMC Bioinf 7
Millar C, Grunstein M (2006) Genome-wide patterns of histone modifications in yeast. Nat Rev Mol Cell Biol 7
Noble W, Ben-Hur A (2007) Integrating information for protein function prediction. In: Lengauer T (ed) m genomes to therapies, vol 3, Wiley, New York, pp 1297–1314
O'Connor T, Wryck J (2007) Chromatindb: a database of genome-wide histone modification patterns for saccharomyces cerevisiae. Bioinformatics 23
Pavesi G, Valentini G (2009) Classification of co-expressed genes from dna regulatory regions. Information Fusion 10
Pavlidis P et al (2002) Learning gene functional classification from multiple data. J Comput Biol 9
Rosset S et al (2004) Boosting as a regularized path to a maximum margin classifier. J Mach Learn Res 5
Spellman P et al (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomices cerevisiae* by microarray hybridization. Mol Biol Cell 9:3273–3297
Subramanian A et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102
Yuan Y et al (2007) Prediction gene expression from sequence: a reexamination. PLOS Comp Biol 3
Zhu J et al (2004) Multi-class adaboost. Statistics and its Interface 2