

# Data fusion based gene function prediction using ensemble methods

Matteo Re and Giorgio Valentini

D.S.I. - Dipartimento di Scienze dell'Informazione  
Università degli Studi di Milano

March 19, 2009

# Gene function prediction

## Gene function prediction:

Given a list of genes, a set of features describing each gene and a reference functional ontology (i.e. Gene Ontology, the FUNctional CATalogue) the goal is to predict the function of each gene.

The first gene function prediction experiments were all based on the use of a **single** source of information. But ...

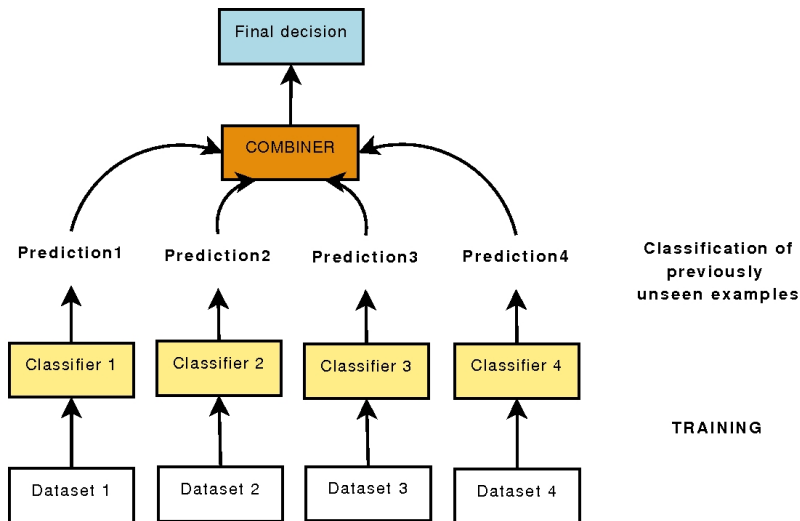
- There are many sources of information that could be predictive of gene function.
- The number of the publicly available biomolecular datasets is constantly growing in the last years as effect of recent advances in high throughput biotechnologies.

# Heterogeneous biomolecular data integration

## Strategies proposed in literature:

- **Vector-space integration:** the vectors describing the same set of genes in different datasources are concatenated and then feed to a **single** classifier [4].
- **Kernel Fusion methods:** Different kernel matrices, each representing the same set of genes in different datasets, are fused using various techniques and then the resulting "integrated" kernel matrix is used to train the final classifier [3].
- **Graphical models:** They provides a probabilistic framework for data integration. Modeling is achieved by representing local probabilistic dependencies. Are often based on Bayesian methods [5].
- **Networks integration:** This approach aims to integrate several networks of functional relationships into a single network [2].

# Heterogeneous data integration: the ensemble system approach



# Reasons for ensemble systems in data fusion based gene function prediction:

- Structurally different datasets can be easily integrated because the fusion is performed at **decision level** (in the intermediate feature space).
- As new datasets (or updates of existing ones) are made available ensemble systems are able to embed the new data (or to update the existing ones) simply by retraining **only the classifiers devoted to these datasets** without retraining the entire system.
- Ensembles of classifiers **scale well** with the number of the available datasources.

# Choice of the combination strategy: (I)

- **Categorical output**: the most commonly adopted combination strategy is the **majority voting**.
- **Continuous valued output**: the most adopted strategy is the **weighted averaging**. In this approach the final support for the appartenance of the instance  $\mathbf{x}$  in a learning problem involving  $C$  classes and  $T$  classifiers is calculated as:

$$\mu_j(\mathbf{x}) = \sum_{t=1}^T w_t D_{t,j}(\mathbf{x}) \quad \text{where } j \in \{1, 2, \dots, C\}$$

the weights could be computed using a convex combination rule ( $w_t^c$ ) or a logarithmic transformation ( $w_t^{\log}$ ):

$$w_t^c = \frac{F_t}{\sum_{t=1}^T F_t} \quad w_t^{\log} \propto \log \frac{F_t}{1 - F_t} \quad (1)$$

## Choice of the combination strategy: (II)

In a classification problem with  $T$  base learners and  $C$  classes:

Let  $DP(x)$  be a matrix composed by the  $d_{t,j}$  elements representing the support given by the  $t^{th}$  classifier to the appartenance of  $x$  to a class  $w_j$ .

Call this matrix a **Decision profile**.

Let  $DT_j$  be the averaged decision profile obtained from  $\mathbf{X}_j$ , the set of training instances belonging to the class  $w_j$ . Call this matrix **Decision Templates** [6].

$$DT_j = \frac{1}{|\mathbf{X}_j|} \sum_{\mathbf{x} \in \mathbf{X}_j} DP(\mathbf{x}) \quad (2)$$

The similarity  $S$  between the decision template  $DT_j$  for a class  $w_j$ ,  $1 \leq j \leq C$ , and the decision profile for a given test instance  $\mathbf{x}$  is:

$$S_j(\mathbf{x}) = 1 - \frac{1}{T \times C} \sum_{t=1}^T \sum_{k=1}^C [DT_j(t, k) - d_{t,k}(\mathbf{x})]^2 \quad (3)$$

# Ensemble selection:

The '**Test and Select**' [1] method allow the selection of subsets of classifiers during the construction of an ensemble system.

## Modified version:

- 1 Separately for each available dataset, selection of the most significant features (two sample t-test with BH correction for multiple test).
- 2 Training of the component classifiers on the heterogeneous data sources each with feature subsets selected at point 1.
- 3 Ranking of the  $n$  learners according to the F-measures collected during "internal" cross-validation on the training set.
- 4 Evaluation of the ensembles formed by the best 2,3 and 4 component learners.

# Experimental setup (I): datasets

Code	Dataset	examples	features	description
D1	Protein domain binary	3529	4950	protein domains obtained from <i>Pfam</i> database [9]
D2	Protein domain log-E	3529	5724	<i>Pfam</i> protein domains with log E-values computed by the <i>HMMER</i> software toolkit
D3	Gene expression	4532	250	merged data of Spellman [11] and Gasch [10] experiments
D4	PPI - BioGRID	4531	5367	protein-protein interaction data from the <i>BioGRID</i> [7] database
D5	PPI - STRING	2338	2559	protein-protein interaction data from the <i>STRING</i> [8]
D6	Pairwise similarity	3527	6349	Smith and Waterman log-Evalues between all pairs of yeast sequences

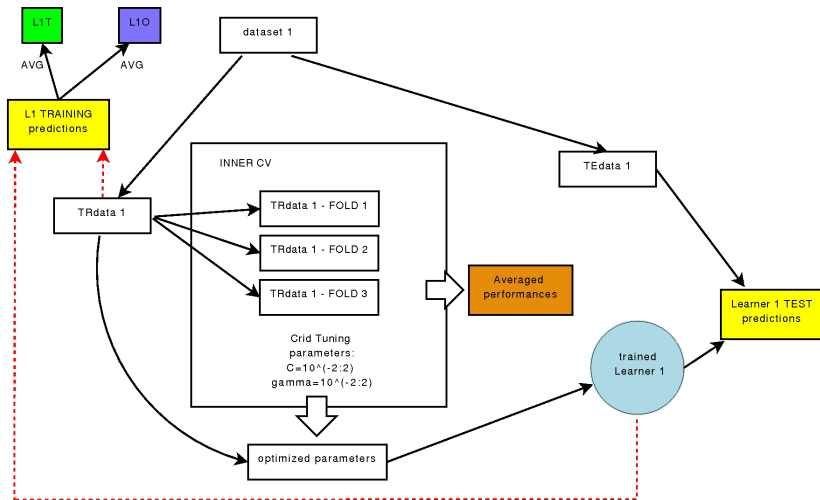
The datasets were merged by **intersection** resulting into a final collection of **1910** genes.

# Experimental setup (II): Functional labelling (MIPS FUNCAT [12])

Code	Description	Code	Description
01	METABOLISM	20	CELLULAR TRANSPORT AND TRANSPORT ROUTES
02	ENERGY	30	CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM
10	CELL CYCLE AND DNA PROCESSING	32	CELL RESCUE, DEFENSE AND VIRULENCE
11	TRANSCRIPTION	34	INTERACTION WITH THE ENVIRONMENT
12	PROTEIN SYNTHESIS	40	CELL FATE
14	PROTEIN FATE		
16	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)	42	BIOGENESYS OF CELLULAR COMPONENTS
18	REGULATION OF METABOLISM AND PROTEIN FUNCTION	43	CELL TYPE DIFFERENTIATION

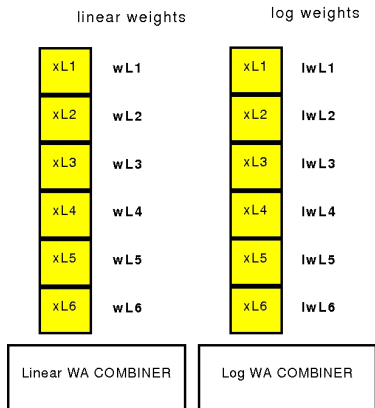
The entire experiment was splitted into **15** independent **binary** classification tasks.

# Experimental setup (II): classifiers and ensemble training



# Experimental setup (III): combining classifier outputs

for any TEST instance  $x$  :



DT(target)

T O

L1T	1-L1T
L2T	1-L2T
L3T	1-L3T
L4T	1-L4T
L5T	1-L5T
L6T	1-L6T

DT(other)

T O

L1O	1-L1O
L2O	1-L2O
L3O	1-L3O
L4O	1-L4O
L5O	1-L5O
L6O	1-L6O

DP(x)

T O

xL1	1-xL1
xL2	1-xL2
xL3	1-xL3
xL4	1-xL4
xL5	1-xL5
xL6	1-xL6

Similarity

Similarity

Decision Templates  
COMBINER

# Results: averaged performances

using all the base learner					
Metric	$L_{best}$	$L_{avg}$	$E_{lin}$	$E_{log}$	$E_{DT}$
F	0.4816	0.3470	0.4403	0.4112	0.5302
rec	0.3970	0.2859	0.3304	0.2974	0.4446
prec	0.6785	0.5823	0.8179	0.8443	0.7034
spec	0.9516	0.9533	0.9798	0.9850	0.9594

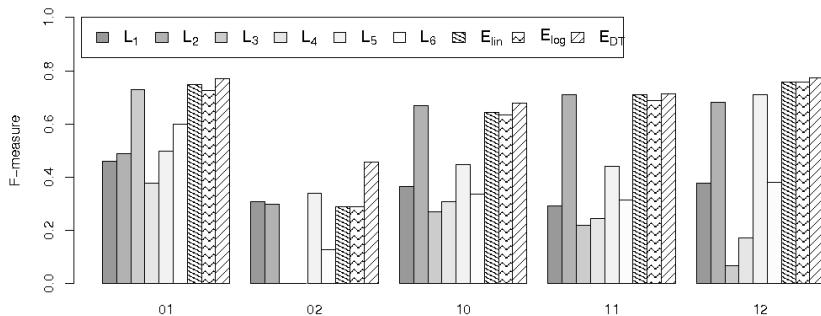
  

+ test and select					
Metric	$L_{best}$	$L_{avg}$	$E_{lin}$	$E_{log}$	$E_{DT}$
F	0.4816	0.3470	0.5436	0.5441	0.5698
rec	0.3970	0.2859	0.4793	0.4778	0.5164
prec	0.6785	0.5823	0.6723	0.6591	0.6435
spec	0.9516	0.9533	0.9538	0.9573	0.9447

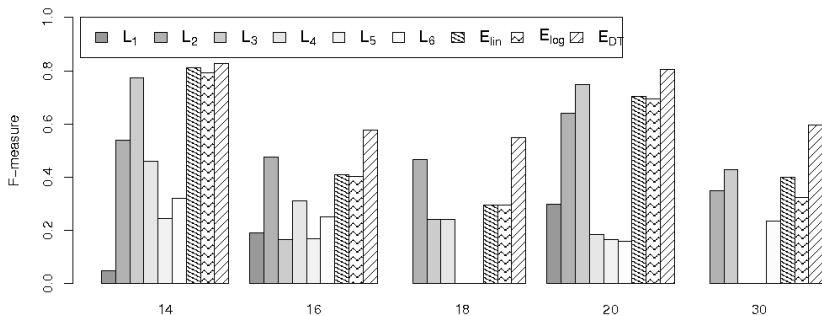
  

+ feature filtering					
Metric	$L_{best}$	$L_{avg}$	$E_{lin}$	$E_{log}$	$E_{DT}$
F	0.4893	0.2638	0.5175	0.4912	0.6310
rec	0.3841	0.1927	0.3987	0.3711	0.5667
prec	0.7278	0.6141	0.8708	0.9042	0.7439
spec	0.9639	0.9775	0.9841	0.9871	0.9552

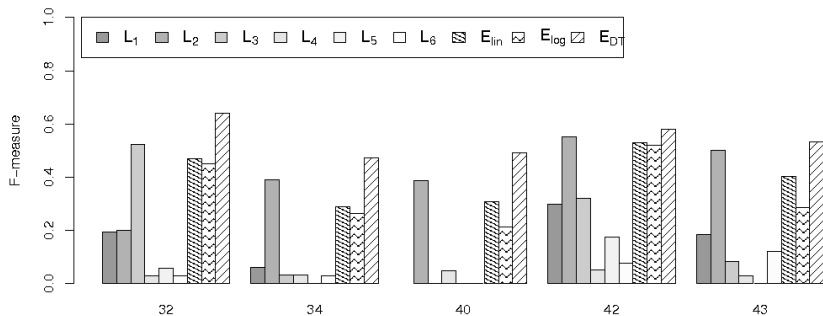
# Results feat. filtering + classifiers selection ( part I )



# Results ( part II )



# Results ( part III )



# Conclusions:

According to the collected F-measures:

- The performances averaged across all the learning tasks are increased by the basic ensemble-based data fusion approach.
- The application of the classifier selection scheme resulted into an additional increment in the performances obtained by all the tested ensemble systems.
- The introduction of the feature filtering step resulted into a decrement in performances of the  $E_{lin}$  and  $E_{log}$  and into an additional increment in performances of the  $DT$  combiner.

We conclude that data fusion realized by mean of ensemble systems is a valuable research line in gene function prediction and Decision Templates may represent a good choice for biomolecular data integration.

# Bibliography I



Sharkey, A., Sharkey N.E., Gerecke, U., Chandroth, G.O.:  
The “Test and Select” Approach to Ensemble Combination  
MCS 2000, Vol. 1857 of LNCS, Springer (2000) 30–44



Chua, H.N., Sung, W.K., Wong L.:  
An efficient strategy for extensive integration of diverse biological data for protein function prediction.  
Bioinformatics  
Oxfordjournals (2007)



Lanckriet, G., De Bie, T., Cristianini, N., Jordan, M., Noble, W.:  
A statistical framework for genomic data fusion.  
Bioinformatics 20 (2004) 2626–2635



Pavlidis, P., Weston, J., Cai, J., Noble, W.:  
Learning gene functional classification from multiple data.  
J. Comput. Biol. 9 (2002) 401–411



Guan, Y., et al.:  
Predicting gene function in a hierarchical context with an ensemble of classifiers.  
Genome Biology 9 (2008)



Kuncheva, L., Bezdek, J., Duin, R.:  
Decision templates for multiple classifier fusion: an experimental comparison.  
Pattern Recognition 34 (2001) 299–314

# Bibliography II



Stark, C., et al.:

BioGRID: a general repository for interaction datasets.  
Nucl. Acids Res. **34** (2006) D535–D539



vonMering, C., et al.:

STRING: a database of predicted functional associations between proteins.  
Nucl. Acids Res. **31** (2003) 258–261



Finn, R., et al.:

The Pfam protein families database.  
Nucl. Acids Res. **36** (2008) D281–D288



Gasch, P., et al.:

Genomic expression programs in the response of yeast cells to environmental changes.  
Mol. Biol. Cell **11** (2000) 4241–4257



Spellman, P., et al.:

Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.  
Mol. Biol. Cell **9** (1998) 3273–3297



Ruepp, A., et al.:

The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes.  
Nucl. Acids Res. **32** (2004) 5539–5545