Matteo Re, Giorgio Valentini
{re,valentini}@dsi.unimi.it

Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction

DSI, Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano

The gene function prediction problem

Is a complex prediction problem characterized by the following features:

1) Each gene can be assigned to **multiple** functional classes (multiclass multilabel classification problem)

2) Functional classes are structured according to a predefined hierarchy (a **DAG** for the Gene Ontology and a **forest of trees** for the MIPS FunCAT)

3) Classes are often **unbalanced** (with negative examples usually exceeding the positive ones)

4) **Multiple sources of information** can be used to predict gene function (problem: **heterogeneous data integration**)

Objective of the experiment

To assess if the performances achievable in data integration based gene function prediction by ensemble systems are comparable/competitive with those of state-of-the-art data integration methods.

To this end:

- we programmatically tested **simple** ensemble methods

- we programmatically used **simple** model tuning techniques

- we **avoided** the use of information about the hierarchical relationships between the terms of the functional ontology (**flat prediction**: 1 binary prediction task for each tested functional term)

Heterogeneous data integration

Existing approaches:

Drawbacks

- **Graphs** and FLN	Karaoz et al. (2004) Chua et al. (2008)	**data type** limitations
- **Vector Space Integration** (VSI)	Pavlidis et al. (2002)	limited **modularity**
- **Kernel Fusion methods**	(SDP) Lanckriet et al.(2004) Lewis et al. (2006)	limited **scalability**

Possible alternatives:

Advantages

- **Ensemble systems**

No data type limitations

High modularity

Good scalability

- Weighted average
- Naive Bayes combination — Titterington et al.(1981)
- Decision Templates — Kuncheva et al.(2001)

Heterogeneous data integration

Ensemble systems performances in heterogeneous data integration for gene function prediction have not been deeply investigated.

Only two papers are dedicated specifically to this topic:

Predicting gene function in a hierarchical context with an ensemble of classifiers.
Guan, Y., Myers, C., Hess, D., Barutcuoglu, Z., Caudy, A., Troyanskaya, O.
Genome Biology 9 (2008)
Naive-Bayes integration of the outputs of SVMs trained with multiple sources of data

Consistent probabilistic output for protein function prediction.
Obozinski, G., Lanckriet, G., Grant, C., M., J., Noble, W.. Genome Biology 9 (2008)
Logistic regression for combining the output of several SVMs trained with different data and kernels in order to produce probabilistic outputs corresponding to specific GO terms

Choice of the model organism

Whole-genome **functional annotation coverage**:

Species	genes with experimental annotation	total annotated genes	%of genes with at least 1 experim. annotation	# of genes
Yeast	4947	5794	**85.4** %	5794
Mouse	10621	18386	**57.8** %	27289

Use and misuse of the gene ontology annotations
Rhee SY et. al. - Nat. Rev. Genetics – Vol. 9 – July 2008

Choice of the model organism

Assessment of **GO annotation quality**:

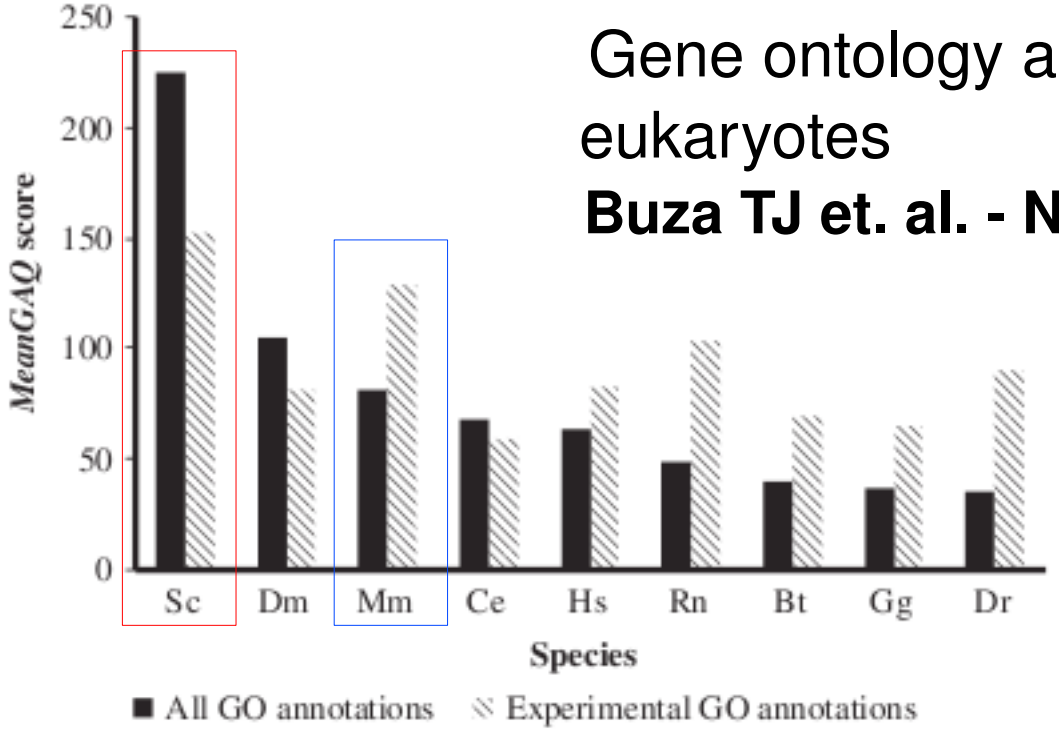
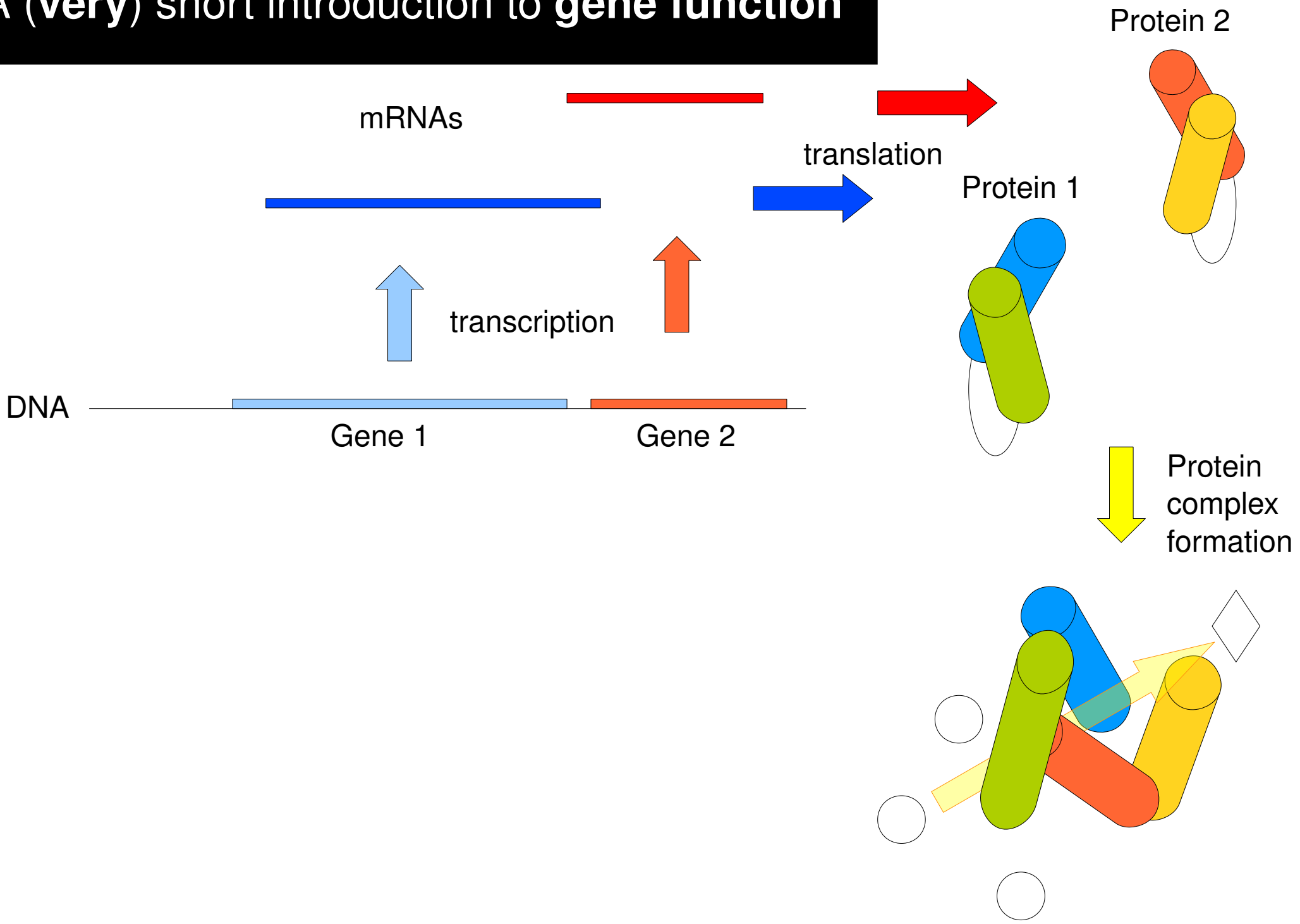


Figure 3. Mean GO Annotation Quality (*GAQ*) scores for each species. To quantify GO annotation quality, we combined annotations (number of annotations per gene product), 'depth' (*Dd*) and evidence quality (*ECR*) to create the GO Annotation Quality (*GAQ*) score. The average *GAQ* score for *S. cerevisiae* (Sc), *D. melanogaster* (Dm), *M. musculus* (Mm), *H. sapiens* (Hs), *C. elegans* (Ce), *R. norvegicus* (Rn), *B. taurus* (Bt), *G. gallus* (Gg) and *D. renio* (Dr) (as at 05/05/2007) is shown. GO annotation founder species have higher overall *meanGAQ* scores than species with more recent GO annotation efforts. Higher scores are found in *Sc*, *Mm*, *Rn* and *Dr*, when computing *meanGAQ* scores from annotations made using only direct experimental evidence codes.

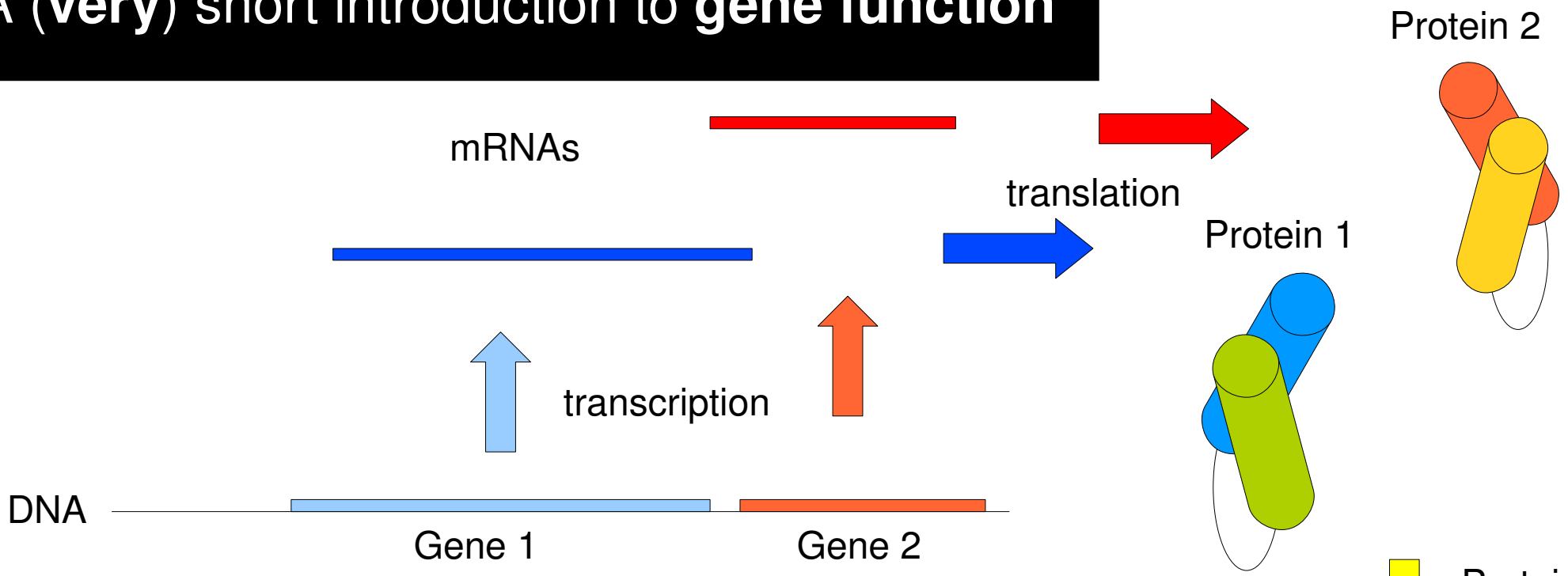
Gene ontology annotation quality analysis in model eukaryotes

Buza TJ et. al. - Nucleic acids research, 2008, 1-9

A (**very**) short introduction to **gene function**



A (**very**) short introduction to **gene function**

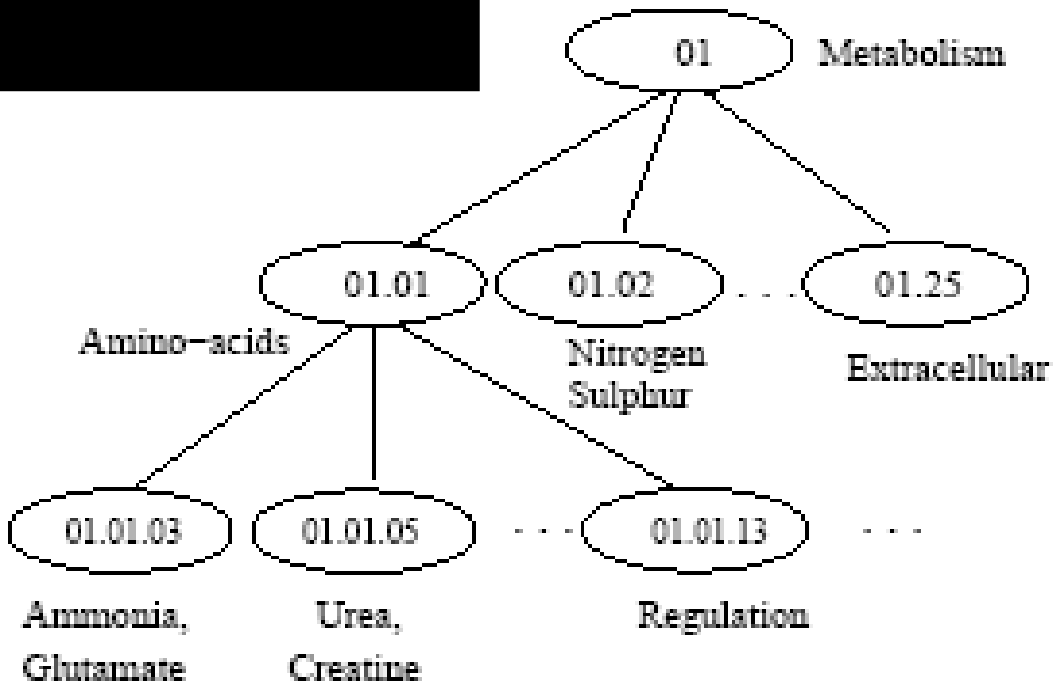


Code	Dataset	examples	features
Dppi1	PPI – STRING	2338	2559
Dppi2	PPI – BioGRID	4531	5367
Dpfam1	Protein domain log-E	3529	5724
Dpfam2	Protein domain binary	3529	4950
Dexpr	Gene expression	4532	250
Dseq	Pairwise similarity	3527	6349

Integration by **intersection** : **1900** yeast genes

Genes **functional labelling**

Functional Annotation	
Gene ID	Class Labels
ytq0045	02.11
ytq0045	20.01.15
ytq0061	02.13.03
...	...

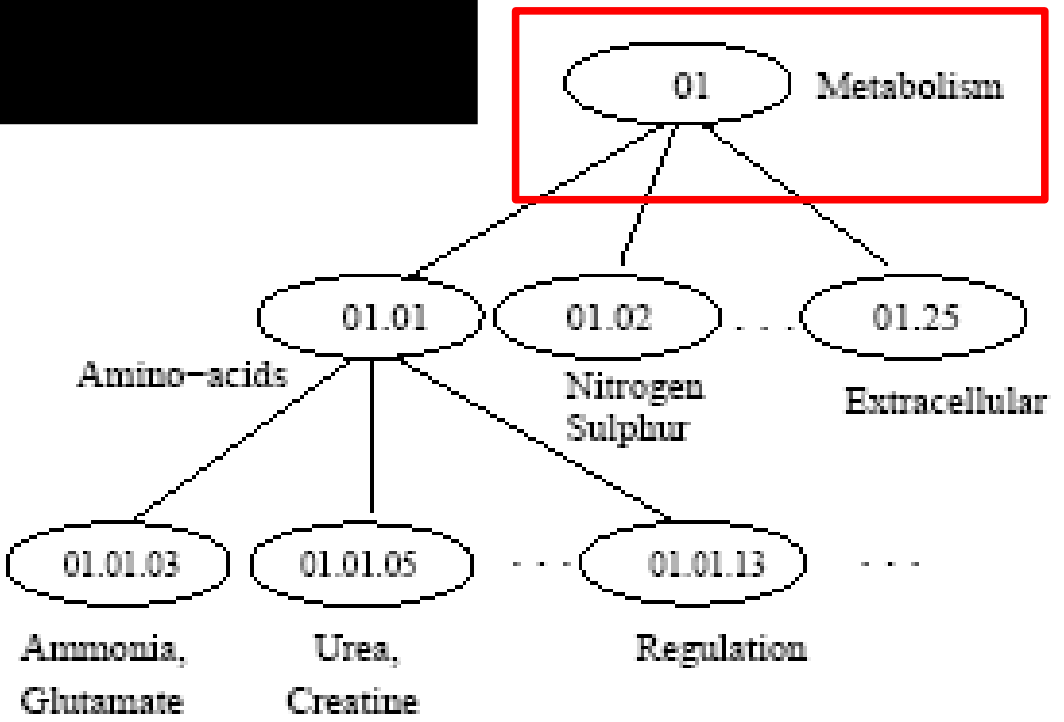


MIPS Functional Catalogue (**FunCAT**)

- FunCat hierarchy - **forest of multiple class trees**
- More than **400** classes
- Each gene can have **multiple** classes

Genes **functional labelling**

Functional Annotation	
Gene ID	Class Labels
ytq0045	02.11
ytq0045	20.01.15
ytq0061	02.13.03
...	...



MIPS Functional Catalogue (**FunCAT**)

Code	Description	Code	Description
01	Metabolism	20	Cellular transport and transport routes
02	Energy	30	Cellular communication/
10	Cell cycle and DNA processing		Signal transduction mechanism
11	Transcription	32	Cell rescue, defense and virulence
12	Protein synthesis	34	Interaction with the environment
14	Protein fate	40	Cell fate
16	Protein with binding function or cofactor requirement	42	Biogenesis of cellular components
18	Regulation of metabolism and protein function	43	Cell type differentiation

Experimental setup

1 vs other FunCAT class prediction

For each FunCAT class:

- normalization of the data sources (mean and sd)
- randomly split the gene set in a training (70%) and a test (30%) set

Vector Space Integration (VSI)

TRAINING

- **concatenation** of the training vectors
- train a probabilistic SVM (**gaussian** kernel) (C and gamma in (10^{-5} to 10^{5}))

TEST:

- **concatenation** of the test vectors
- predict the FunCAT class of test examples

Experimental setup

1 vs other FunCAT class prediction

Kernel Fusion (KF)

For each FunCAT class:

TRAINING:

foreach gamma in (10^{-5} to 10^{5})

foreach data source

- construction of a kernel matrix (gaussian kernel)
- normalization of the kernel matrix (w.r.t. mean and sd)

- sum the kernel matrices
- train a probabilistic SVM, C (fixed value: 10)
- collect performances (Fmeasure)

TEST:

prediction of the test instances using the best performing training model

Experimental setup

1 vs other FunCAT class prediction

Ensemble systems: **component classifiers**

For each FunCAT class:

TRAINING:

- **for each** dataset
  - **train** a *probabilistic* SVM (**gaussian** kernel)
    - *model tuning* based on a 3 fold CV scheme on the training set:
      grid tuning of **C** and **gamma** (both ranging from 10^{-5} to 10^{5}).
      (collect performances: Fmeasure averaged accross the CV folds)
    - train the final models using the tuned parameters on the entire training set
    - predict the training examples (required for the DT and NB combiners)

TEST:

- **predict** the FunCAT class of the test examples (current class or other)

Experimental setup

1 vs other FunCAT class prediction

Ensemble systems:

For each FunCAT class:

TEST: (intermediate feature space constituted by the test set predictions produced by the component classifiers)

- **Weighted average (with linear and logarithmic weights)**
- **Naive bayes combiner**
- **Decision Templates**

all the ensemble methods have been implemented according to the formulations presented in:

Combining patterns classifiers, methods and algorithms
L.I. Kuncheva, Wiley-interscience, 2004

Performances evaluation

Performances **averaged** across **15** classification tasks

Base learners performances:

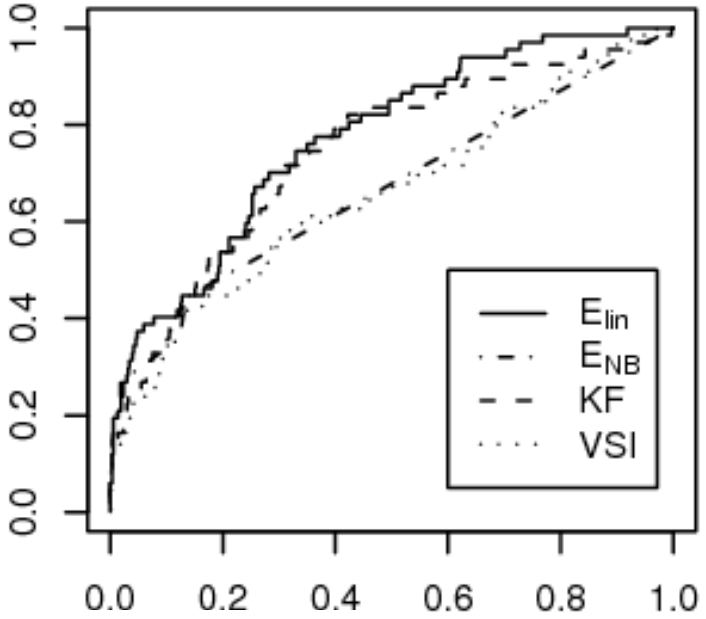
Metric	D_{ppi1}	D_{ppi2}	D_{pfam1}	D_{pfam2}	D_{expr}	D_{seq}
F	0.3655	0.4818	0.2363	0.3391	0.2098	0.4493
rec	0.2716	0.3970	0.1457	0.2417	0.1571	0.5019
prec	0.6157	0.6785	0.7154	0.6752	0.3922	0.4162
AUC	0.7501	0.8170	0.6952	0.6995	0.6507	0.7469

Data integration performances:

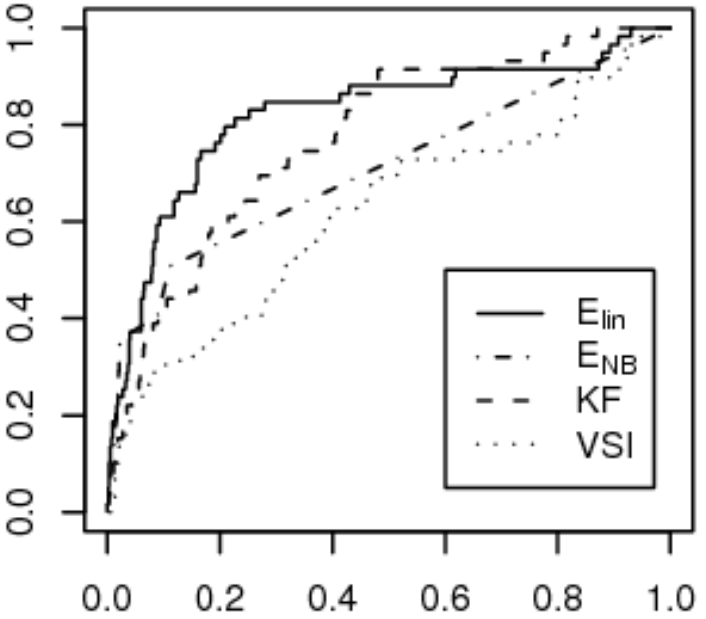
Metric	E_{lin}	E_{log}	E_{dt}	E_{NB}	VSI	KF	D_{avg}	D_{ppi2}
F	0.4347	0.4111	0.5302	0.5174	0.3213	0.3782	0.3544	0.4818
rec	0.3304	0.2974	0.4446	0.6467	0.2260	0.3039	0.2859	0.3970
prec	0.8179	0.8443	0.7034	0.5328	0.6530	0.6293	0.5823	0.6157
AUC	0.8642	0.8653	0.8613	0.7933	0.7238	0.7775	0.7265	0.8170

Performances evaluation: **ROC curves**

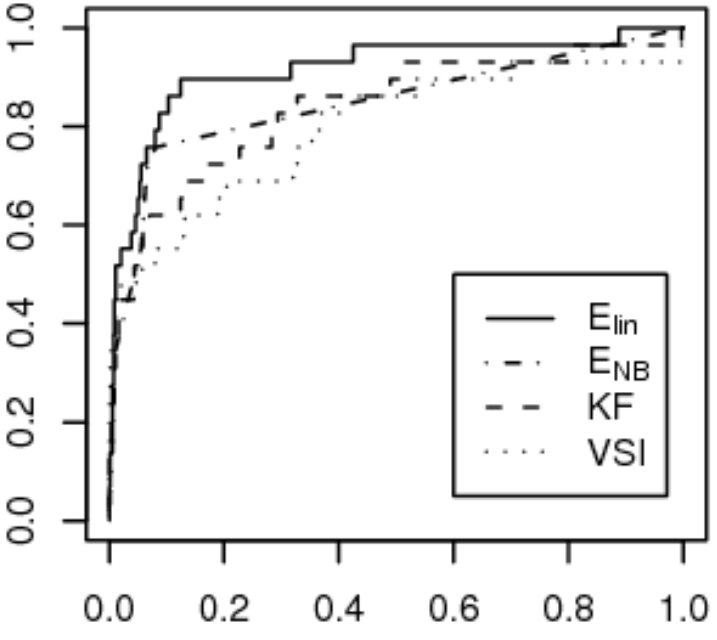
Cell rescue



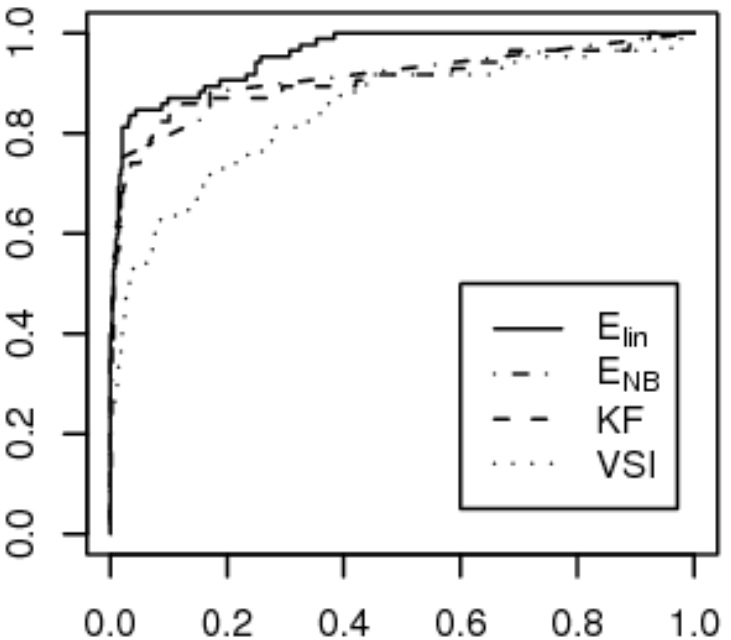
Interaction w. environment



Cellular communication



Protein synthesis



AUCs comparisons

Comparisons of AUCs in the 15 separated classification tasks (non parametric **Mann-Withney statistic,** significance cutoff: **0.01**)

(Results are reported in terms of wins - ties - losses counts)

	VSI	E_{log}	E_{lin}	E_{dt}	E_{NB}
E_{log}	13-2-0	-	-	-	-
E_{lin}	13-2-0	0-14-1	-	-	-
E_{dt}	13-2-0	1-13-1	1-11-3	-	-
E_{NB}	9-6-0	0-2-13	0-2-13	0-2-13	-
KF	3-12-0	0-6-9	0-6-9	0-6-9	0-10-5

	D_{ppi1}	D_{ppi2}	D_{pfam1}	D_{pfam2}	D_{expr}	D_{seq}
E_{lin}	11-4-0	4-11-0	15-0-0	14-1-0	15-0-0	13-2-0
E_{log}	11-4-0	4-11-0	15-0-0	14-1-0	15-0-0	13-2-0
E_{dt}	11-4-0	4-11-0	15-0-0	14-1-0	15-0-0	13-2-0
E_{NB}	5-10-0	2-11-2	9-6-0	8-7-0	12-3-0	7-8-0
VSI	1-11-3	0-8-7	2-11-2	1-14-0	4-11-0	0-12-3
KF	1-14-0	0-9-6	5-10-0	5-10-0	11-4-0	3-12-0

Best performing component classifier (D*ppi2*)

Conclusions and ...

In this preliminary work we evaluated the performances of ensemble systems for **heterogeneous biomolecular data integration** (without exploit information about the structure of the functional ontology).

Our results clearly indicated that ensemble systems are **at least comparable** with state of the art heterogeneous data integration methods.

Considering the increasing growing rate of available biomolecular data, the **modularity and scalability** that characterize ensemble methods can favour an easy update of existing sources of data and an easy integration of new ones.

BUT ...

Future directions

We are aware that this is only a preliminar experiment:
THERE IS A LOT OF ROOM FOR IMPROVEMENT

We need to increase the whole **gene catalog coverage** (we plan to reach a coverage comprised between 75 and 80% of the yeast genes)

We need to test the ensembles on a **larger number of functional terms** (hundreds)

We need to integrate a **larger number of data sources** (15... 20?)

this experiment is currently underway ...

...

Questions?