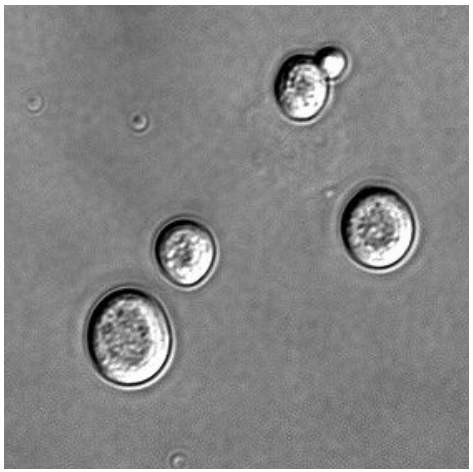
Hierarchical Cost-Sensitive Algorithms for Genome-Wide Gene Function Prediction

Nicolò Cesa-Bianchi

Università degli Studi di Milano



Joint work with: Giorgio Valentini (Milano)

Motivation

- Novel high-throughput biotechnologies generate a wealth of data about genes and gene products
- Manual annotation of gene function becomes infeasible
- For most species the functions of several genes are still unknown or only partially known
- *In silico* methods represent a fundamental tool for gene function prediction at genome-wide and ontology-wide level
- Computational analysis provide predictions that drive the biological validation of gene function

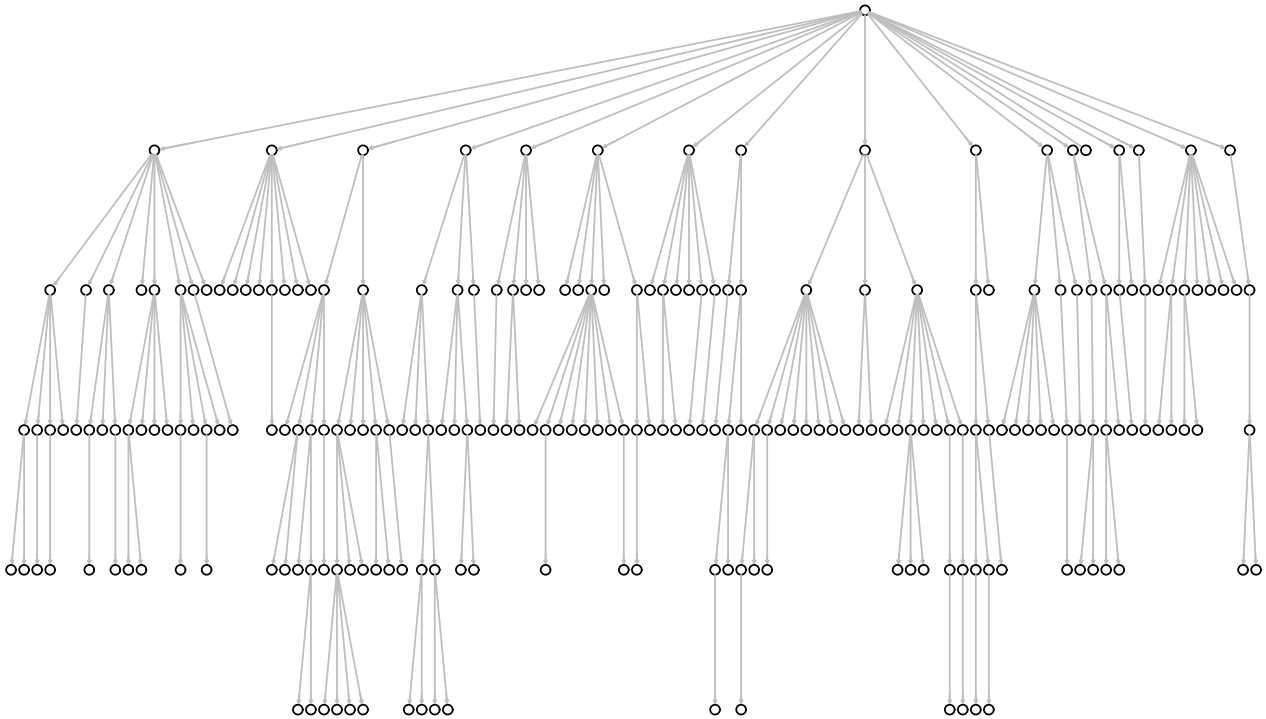
Main issues

Gene function prediction

- Large number of functional classes (in the hundreds for FunCat)
- Taxonomy with multiple functional annotations for each gene
- Sparsity of annotations (more than 80% of genes have less than 4 annotations)

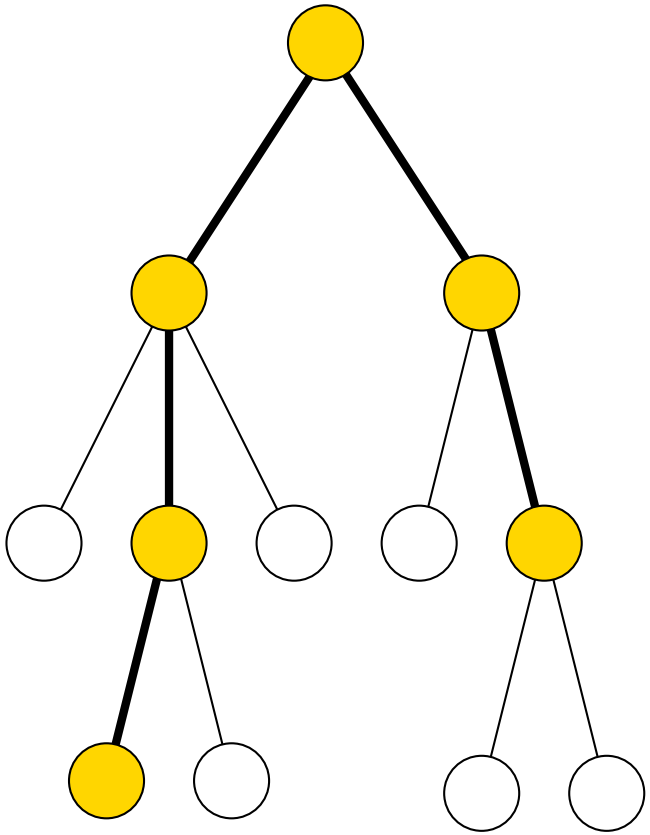
- Uncertainty of annotations
- Multiple sources of data

The hierarchical classification problem



- Genome-wide gene function prediction in *S. cerevisiae*
- About 200 FunCat classes (5 hierarchical levels)
- About 6000 genes

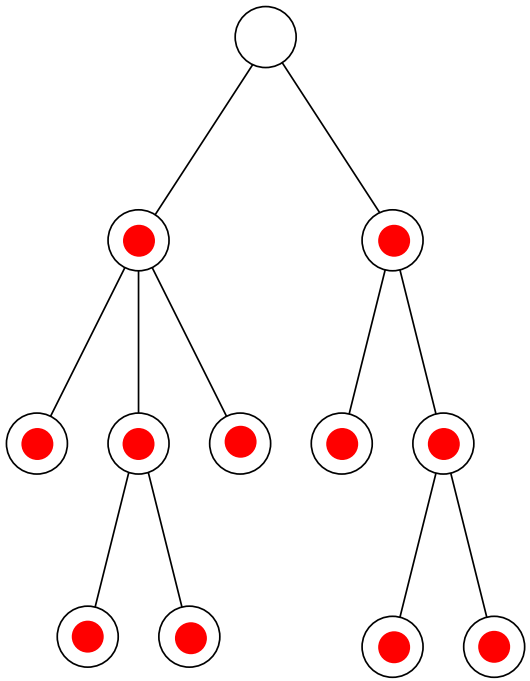
Hierarchical multilabels



True multilabel for gene $\mathbf{x}$

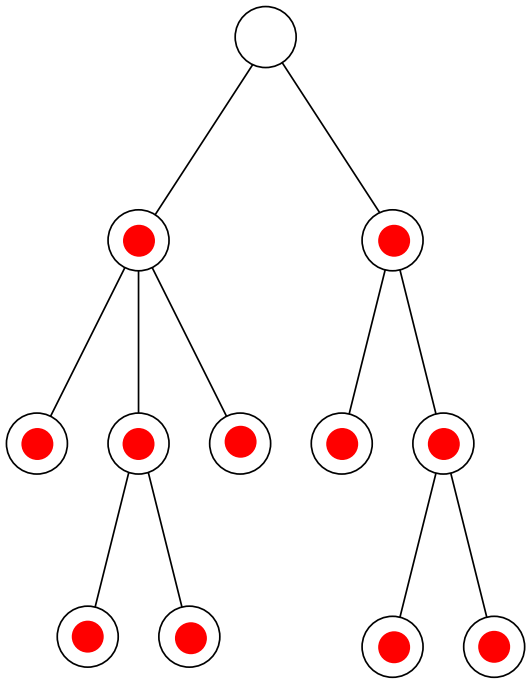
- True path rule
- Multiple paths
- Partial paths

Ensemble methods



- A binary classifier associated to each node of the taxonomy
- Prediction of node i on gene $\mathbf{x}$ is $p_i \in [0,1]$
- We train each node using calibrated SVMs (Gaussian kernels)

Ensemble methods

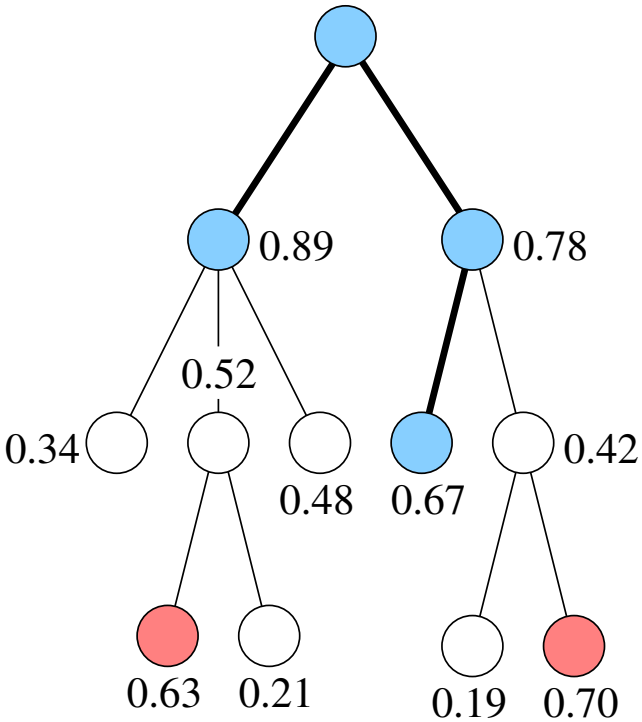


- A binary classifier associated to each node of the taxonomy
- Prediction of node i on gene $\mathbf{x}$ is $p_i \in [0,1]$
- We train each node using calibrated SVMs (Gaussian kernels)

Basic problem:

Given node predictions $p_1, \ldots, p_N$ for gene $\mathbf{x}$ derive the "correct" multilabel $(y_1, \ldots, y_N) \in \{0,1\}^N$

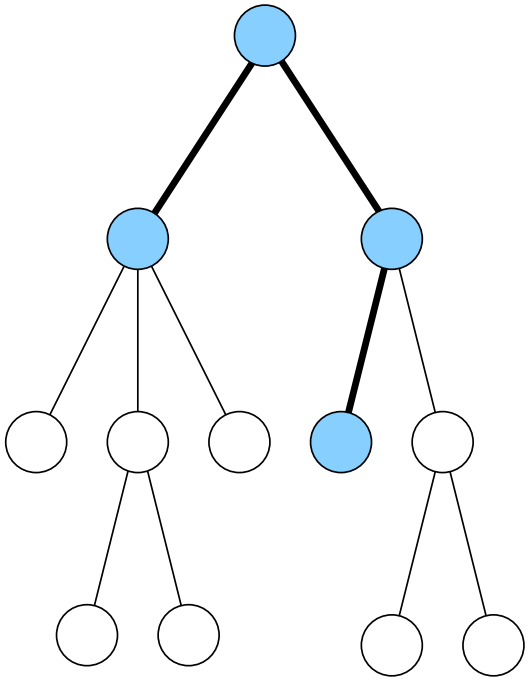
Cost-sensitive hierarchical top-down



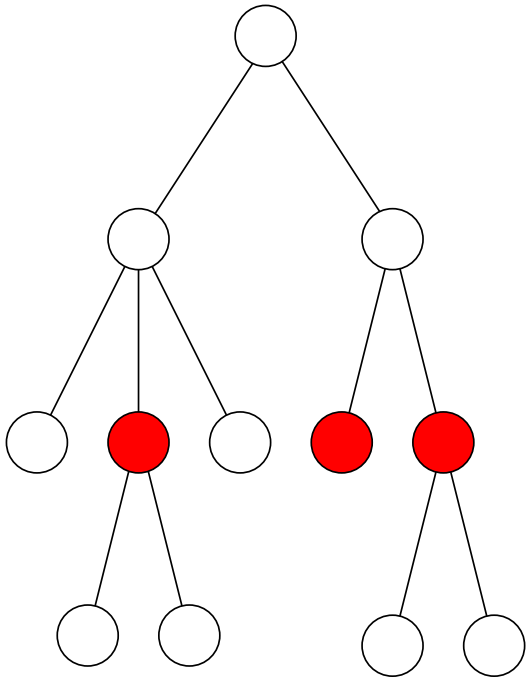
$\tau = 0.6$

- Node i is assigned label $+1$ iff $p_i \geqslant \tau$ (cost-sensitiveness par.)
- Any node violating TPR is then set to -1
- τ decided via cross-validation

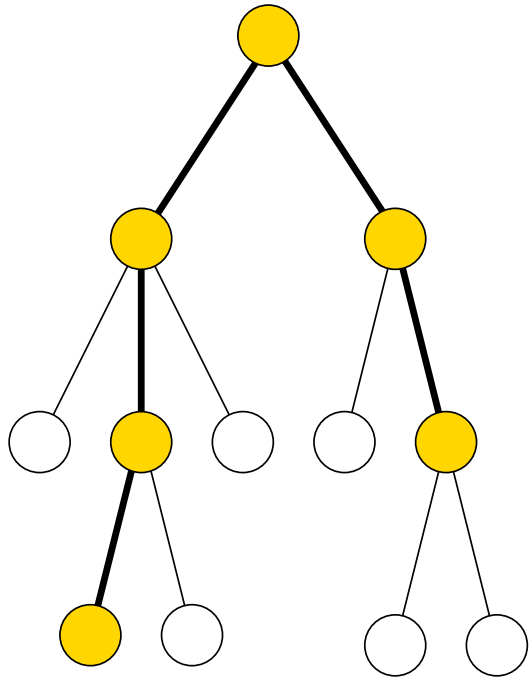
H-loss



Guessed multilabel for gene $\mathbf{x}$



H-loss $= 3$



True multilabel for gene $\mathbf{x}$

Hierarchical Bayesian Algorithm

- $\ell_H(\mathbf{y}, \boldsymbol{\nu})$ is H-loss of guessed multilabel $\mathbf{y} \in \{0,1\}^N$ w.r.t. true multilabel $\boldsymbol{\nu} \in \{0,1\}^N$
- Given node predictions $p_1, \ldots, p_N$ for a gene, HBAYES-CS predicts

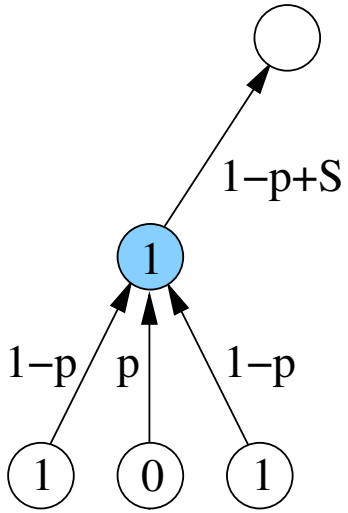
$$\widehat{\mathbf{y}} = \operatorname*{argmin}_{\mathbf{y} \in \{0,1\}^N} \mathbb{E}\left[\ell_H(\mathbf{y}, \mathbf{V})\right]$$

- $\mathbf{V} \in \{0,1\}^N$ is a random vector with law

$$\mathbb{P}\left(V_i = 1 \mid V_{\mathrm{par}(i)} = \nu\right) = \begin{cases} p_i & \text{if} \quad \nu = 1 \\ 0 & \text{if} \quad \nu = 0 \end{cases}$$

- $\widehat{\mathbf{y}}$ is Bayes optimal for H-loss given $p_1, \ldots, p_N$
- $\widehat{\mathbf{y}}$ is computed in linear time via a message-passing algorithm

The Message Passing Algorithm

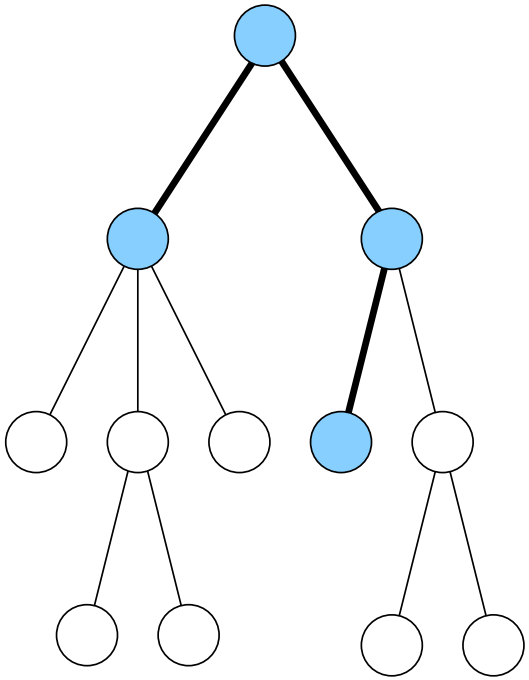


- Each leaf node i is assigned 1 iff $p_i \geqslant \frac{1}{2}$
- Each node i sends to its parent the expected H-loss of its subtree

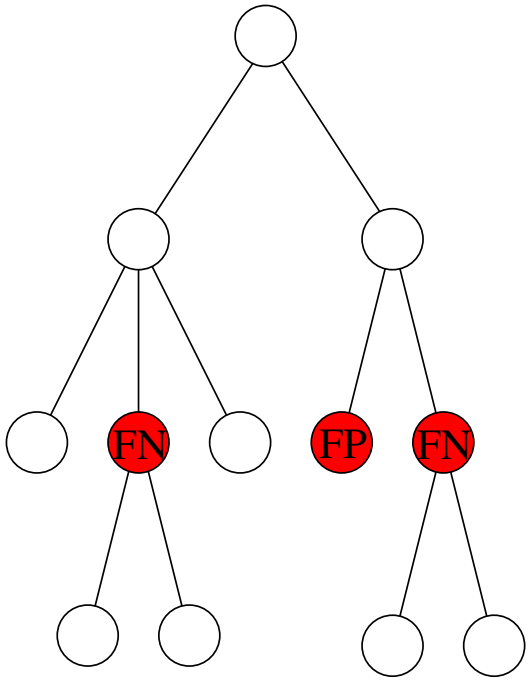
Label assigned to node i

$$\widehat{y}_i = \operatorname*{argmin}_{y \in \{0,1\}} \left(\underbrace{p_i(1-y)}_{\text{Exp. subtree loss if } y=0} + \underbrace{(1-p_i)y + p_i\, y \sum_{j \in \text{child}(i)} S_j}_{\text{Exp. subtree loss if } y=1} \right)$$

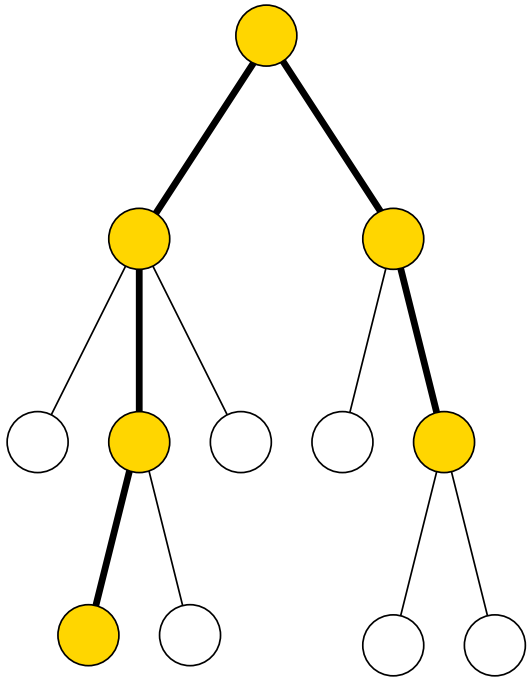
Cost-sensitive H-loss



Guessed multilabel for gene $\mathbf{x}$

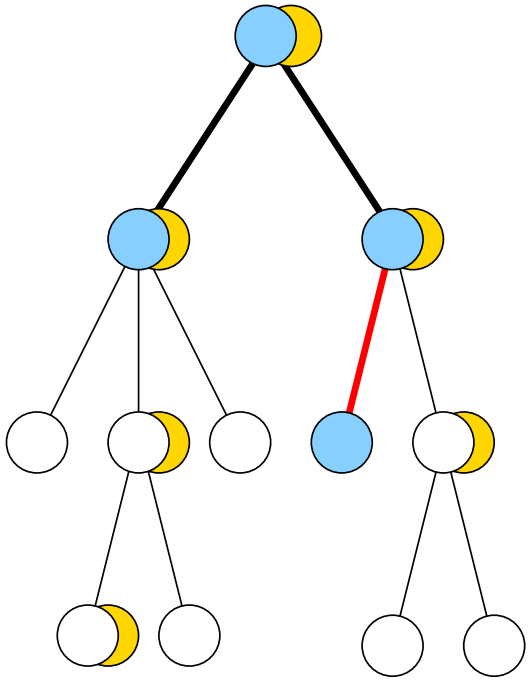


$2 \times \mathrm{FN} + 1 \times \mathrm{FP} = 2\alpha + 1$

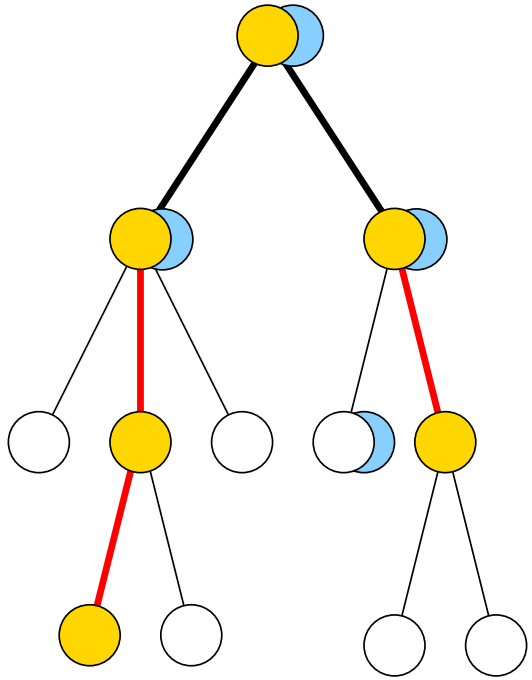


True multilabel for gene $\mathbf{x}$

Hierarchical precision and recall



Precision $= \frac{1}{2}\left(1 + \frac{1}{2}\right)$



Recall $= \frac{1}{2}\left(\frac{1}{3} + \frac{1}{2}\right)$

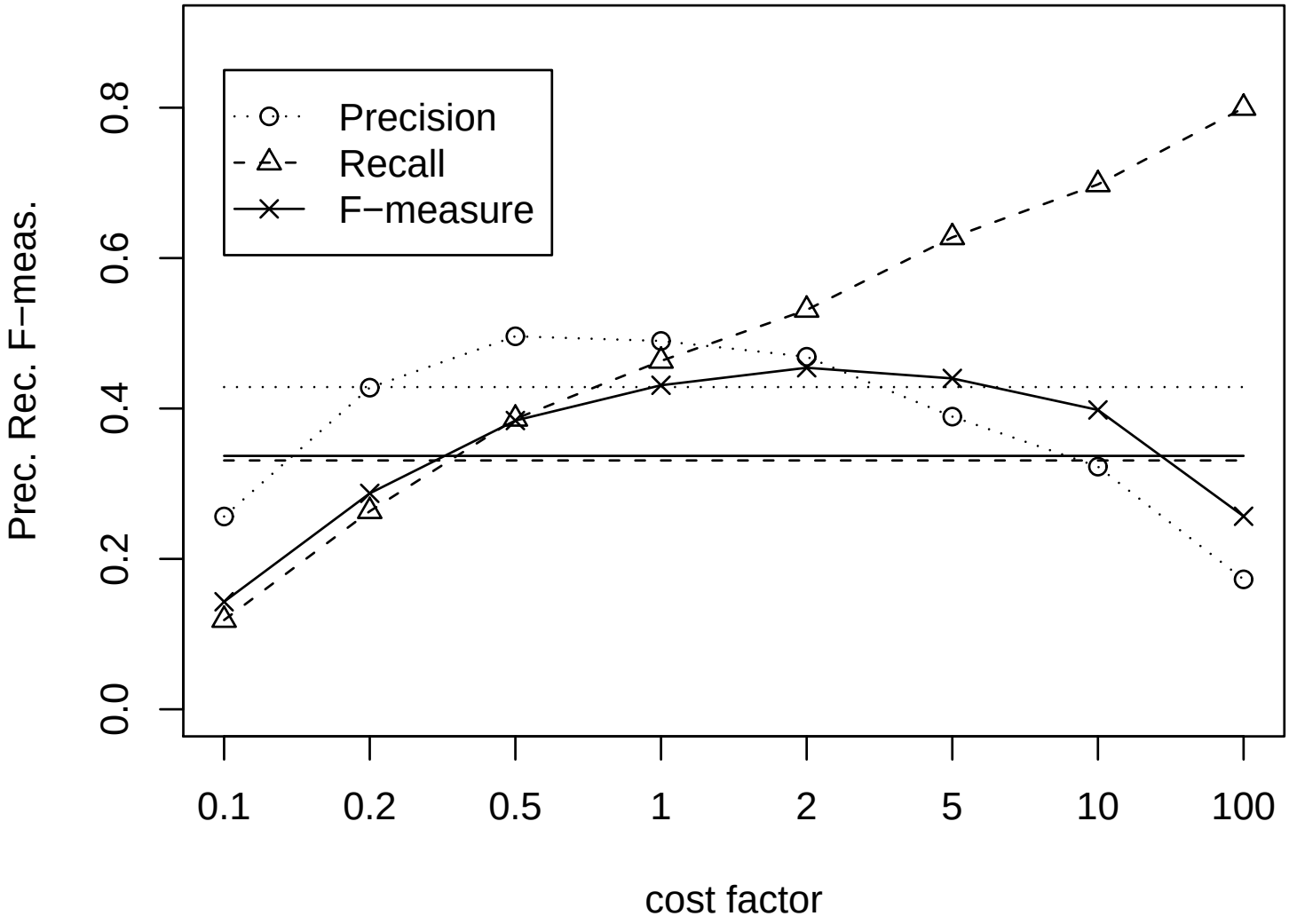
Experimental setup – Features sets

- **Pfam** Presence/absence of protein domains from *Pfam*
- **Pfam-2** Pfam with log E-values (HMMER)
- **Phylo** log E-values (BLAST) against the genome in 24 organisms
- **Expr** Gene expression data (77 conditions) + transcriptional response to env. stress (173 conditions)
- **PPI-BG, PPI-VM** protein-protein interaction data (BioGRID *et al.*)
- **SP-sim** log E-values of pairwise similarities between yeast genes

Experimental setup – Features sets

Data set	genes	features	classes
Pfam-1	3529	4950	211
Pfam-2	3529	5724	211
Phylo	2445	24	187
Expr	4532	250	230
PPI-BG	4531	5367	232
PPI-VM	2338	2559	177
SP-sim	3527	6349	211

Sensitivity to cost factor in HBAYES-CS



Results – Hierarchical F-measure

Methods	**Data sets**							
	Pfam-1	Pfam-2	Phylo	Expr	BG	VM	sim	ALL
HTD	0.38	0.01	0.25	0.23	0.15	0.42	0.34	0.25
HTD-CS	0.42	0.20	0.30	0.26	0.31	0.46	0.42	0.34
HBAYES-CS	0.45	0.20	0.27	0.26	0.29	0.43	0.45	0.34

win-tie-loss		
Methods	HTD-CS	HTD
HBAYES-CS	2-4-1	6-1-0
HTD-CS	-	7-0-0

Per-level performance – Macroaveraged F-measure

Pfam Protein domain									
Lev.	HTD			HTD-CS			HBAYES-CS		
	P	R	F	P	R	F	P	R	F
1	0.76	0.31	0.43	0.66	0.37	0.47	0.74	0.35	0.47
2	0.69	0.29	0.39	0.61	0.35	0.43	0.65	0.33	0.43
3	0.62	0.25	0.35	0.55	0.30	0.38	0.58	0.30	0.38
4	0.56	0.23	0.31	0.53	0.27	0.35	0.54	0.27	0.34
5	0.47	0.20	0.27	0.46	0.22	0.29	0.45	0.20	0.26

Conclusions

- Sparse multilabels call for:
  1. Cost-sensitive methods
  2. Hierarchical precision and recall
- Complex methods (HBAYES-CS) perform as simpler ones (HTD-CS)
  $\rightarrow$ Overfit of noisy annotations at lower levels?
- No method is actually designed to optimize true performance measure
- In progress:
  1. Asymmetric kernels to compute hierarchical precision and recall
  2. Data integration combined with hierarchical classification