Functional inference in FunCat through the combination of hierarchical ensembles with data fusion methods

Matteo Re

e-mail: re@dsi.unimi.it

DSI - Dipartimento di Scienze dell’Informazione
Università degli Studi di Milano

joint work with:
Nicolò Cesa-Bianchi
Giorgio Valentini
Genome and ontology-wide gene function prediction

- Novel high-throughput biotechnologies accumulated a wealth of data about genes and gene products.
- Manual annotation of gene function is time consuming and expensive and becomes infeasible for growing amount of data.
- For most species the functions of several genes are unknown or only partially known: “in silico” methods represent a fundamental tool for gene function prediction at genome-wide and ontology-wide level [Friedberg, 2006].
- Computational analysis provide predictions that can be considered hypotheses to drive the biological validation of gene function [Pena-Castillo et al. 2008].
Main characteristics of the GFP problem

- Large number of functional classes: hundreds (*FunCat*) or thousands (*Gene Ontology (GO)*).
- Multiple annotations for each gene (multilabel classification)
- Hierarchical relationships between functional classes (tree forest for *FunCat*, direct acyclic graph for GO)
- Class frequencies are unbalanced: positive examples are usually largely lower than negatives
- Multiple sources of data available: each type captures specific functional characteristics of genes/gene products
The “true path rule”

“An annotation for a class in the hierarchy is automatically transferred to its ancestors, while genes unannotated for a class cannot be annotated for its descendants”.
A single type of data **is not sufficient** to predict **all** the possible gene functions!
Introduction

Genome and ontology-wide gene function prediction

Data Fusion methods

**Weighted averaging:**  
\[
\hat{P}(V_i = 1 \mid g) = \frac{1}{\sum_{s=1}^{L} F_s} \sum_{t=1}^{L} F_t \hat{p}_{t,i}(g)
\]

**Kernel Fusion:**  
\[
K_{\text{ave}}(g, g') = \frac{1}{L} \sum_{t=1}^{L} K_t(x_t, x'_t)
\]

where:

- \( V_i \in \{0, 1\} \): random variable that models the labeling of a gene \( g \) for the class \( \omega_i \in \Omega \)
- \( L \) different sources of biomolecular data \( D_t \), for \( t = 1, \ldots, L \)
- \( \hat{p}_{t,i}(g) \): classifier’s estimate of the probability that \( g \) belongs to \( \omega_i \) using data \( D_t \)
- \( F_t \) is the F-measure assessed on the training data for the \( t \)-th base learner
- \( g \) and \( g' \): a pair of genes, and \( x_t, x'_t \in D_t \) their corresponding pairs of feature vectors.

**WARNING:** there are still inconsistencies w.r.t. the TPR.
How to avoid TPR inconsistencies in the predicted multilabels?

A possible solution is to combine independent local predictions at each functional node in order to obtain multilabels that respects the TPR. This approach has been recently investigated in [Obozinski 2008]. But:

- This work focuses only on the comparison of hierarchical multilabel methods
- Does not take into account the impact of the concurrent use of data fusion and hierarchical multilabel methods
- Does not take into account the potential benefits introduced by the application of cost-sensitive techniques
- It is based on mouse data...
The quality of the functional annotation in mouse are lower than the ones available in yeast.
AIM:

*Evaluation of a two-steps strategy:*

1. For each term of the taxonomy, train a classifier using multiple sources of data
2. Combine the predictions at each node to obtain the multi-label predictions according to an hierarchical ensemble method that takes as input the local estimates probability.

Evaluation of the impact of cost-sensitive methods in hierarchical multilabel GFP.
Hierarchical multilabel ensembles:

A binary classifier associated with each node

Prediction of node i on gene x is $p_i \in [0,1]$.

Basic problem:

Given node predictions $p_1, \ldots, p_N$ for gene x derive the “correct” multilabel $(y_1, \ldots, y_N) \in \{0, 1\}^N$. 
Hierarchical Top-Down (cost sensitive):

Node $i$ is assigned label $+1$ iff $p_i \geq \tau$ AND $y(par(i)) = +1$

Any node violating TPR is then set to $-1$

$\tau = 0.6$  

(cost-sensitivity parameter)
The HBAYES method

A method based on:
- An underlying stochastic model for the multilabels
- The Hierarchical loss (H-loss)
- An approximation of the bayesian-optimal classifier for the H-loss
The H-loss:

The main intuition behind the H-loss:

*if a parent class has been predicted wrongly, then errors in its descendants should not be taken into account.*
Hierarchical Bayesian Algorithm: [Cesa-Bianchi et al. 2006]

- $\ell_H(\hat{y}, v)$ is H-loss for guessed multilabel $\hat{y} \in \{0, 1\}^N$ w.r.t. true multilabel $v \in \{0, 1\}^N$.
- Given node predictions $p_1, \ldots, p_N$ for a gene $x$ HBayes predicts
  \[
  \hat{y} = \arg\min_{y \in \{0, 1\}^N} \mathbb{E}[\ell_H(y, V) \mid x]
  \]
- $V \in \{0, 1\}^N$ is a random vector with law:
  \[
  \mathbb{P}(V = v) = \prod_{i=1}^{N} \mathbb{P}(V_i = v_i \mid V_{\text{par}(i)} = 1, x)
  \]
  for all $v \in \{0, 1\}^N$
  where $\mathbb{P}(V_i = 1 \mid V_{\text{par}(i)} = 0, x) = 0$ for all $i$ and $x$
- $\hat{y}$ is the Bayes optimal for H-loss given $p_1, \ldots, p_N$
The message passing algorithm (I):

Label assigned to node $i$:

$$
\hat{y}_i = \arg\min_{y \in \{0,1\}} \left( c_i (p_i (1 - y)) + \sum_{k \in \text{child}(i)} H_k(\hat{y}) \right)
$$

where $H_k(\hat{y}) = c_k (p_k (1 - \hat{y}_k) + (1 - p_k)\hat{y}_k) + \sum_{j \in \text{child}(k)} H_j(\hat{y})$

Each leaf node is assigned +1 iff $p_i \geq 0.5$

Each node $i$ sends to its parent the expected H-loss of its subtree
Functional inference in FunCat through the combination of hierarchical ensembles with data fusion methods

Hierarchical ensembles

The HBAYES ensemble algorithm

The message passing algorithm (II): CS variant

Each leaf node is assigned +1 iff $p_i \geq 0.5$

Each node i sends to its parent the expected H-loss of its subtree

Label assigned to node $i$:

$$\hat{y}_i = \arg\min_{y \in \{0,1\}} \begin{cases} 

c_i^- (p_i(1-y)) & \text{Exp. subtree loss if } y = 0 \\

+c_i^+((1-p_i)y + p_iy \sum_{k \in \text{child}(i)} H_k(\hat{y})) & \text{Exp. subtree loss if } y = 1
\end{cases}$$

$$c_i^- = \alpha c_i^+ \text{ while keeping } c_i^- + c_i^+ = 2c_i$$
Combination of hierarchical ensembles with data fusion methods

**Experimental setting:**

1. About 2000 genes and 169 FunCat functional classes
2. 6 datasets (PPI, expression levels, protein domains, sequence similarity)
3. 2 Data integration methods, 2 hierarchical multilabel techniques (both in vanilla and cost-sensitive variants) + hierarchical “flat”

**Two-levels of improvements:**

- Improvement of flat predictions through the bottom-up Bayesian and hierarchical top-down correction
- Improvement of the single-source predictions by exploiting multiple sources of data
## Results: Impact of data fusion on flat and hierarchical methods (average F-scores)

<table>
<thead>
<tr>
<th>METHODS</th>
<th>FLAT</th>
<th>HTD</th>
<th>HTD-CS</th>
<th>HB</th>
<th>HB-CS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SINGLE-SOURCE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIOGRID</td>
<td>0.2643</td>
<td>0.3759</td>
<td>0.4160</td>
<td>0.3385</td>
<td>0.4183</td>
</tr>
<tr>
<td>STRING</td>
<td>0.2203</td>
<td>0.2677</td>
<td>0.3135</td>
<td>0.2138</td>
<td>0.3007</td>
</tr>
<tr>
<td>PFAM BINARY</td>
<td>0.1756</td>
<td>0.2003</td>
<td>0.2482</td>
<td>0.1468</td>
<td>0.2395</td>
</tr>
<tr>
<td>PFAM LOGE</td>
<td>0.2044</td>
<td>0.1567</td>
<td>0.2541</td>
<td>0.0997</td>
<td>0.2500</td>
</tr>
<tr>
<td>EXPR.</td>
<td>0.1884</td>
<td>0.2506</td>
<td>0.2889</td>
<td>0.2006</td>
<td>0.2781</td>
</tr>
<tr>
<td>SEQ. SIM.</td>
<td>0.1870</td>
<td>0.2532</td>
<td>0.2899</td>
<td>0.2017</td>
<td>0.2825</td>
</tr>
<tr>
<td><strong>MULTI-SOURCE (DATA FUSION)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KERNEL FUSION</td>
<td><strong>0.3220</strong></td>
<td><strong>0.5401</strong></td>
<td><strong>0.5492</strong></td>
<td><strong>0.5181</strong></td>
<td><strong>0.5505</strong></td>
</tr>
<tr>
<td>WEIGH. VOTING</td>
<td>0.2754</td>
<td>0.2792</td>
<td>0.3974</td>
<td>0.1491</td>
<td>0.3532</td>
</tr>
</tbody>
</table>

- About **2000** genes, **169** FunCat classes and **6** data sources
- 2 data fusion techniques: Kernel Fusion and weighted voting
- Flat, HTD, HBAYES and their cost-sensitive variants
Comparison of F-scores with and without data integration

- Black nodes: better results with data fusion
- White nodes: better results with the best single-source data
- $p$-value $= 2.2 \cdot 10^{-16}$ (Wilcoxon signed-rank sums test)
Synergy between hierarchical, data fusion and cost-sensitive techniques (hierarchical F-score)

<table>
<thead>
<tr>
<th>METHODS</th>
<th>F-SCORE</th>
<th>PREC.</th>
<th>REC.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BI OGRID:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAT</td>
<td>0.1893</td>
<td>0.1253</td>
<td>0.5801</td>
</tr>
<tr>
<td>HTD</td>
<td>0.4311</td>
<td>0.5901</td>
<td>0.3827</td>
</tr>
<tr>
<td>HTD-CS</td>
<td>0.4732</td>
<td>0.5645</td>
<td>0.4650</td>
</tr>
<tr>
<td>HBAYES</td>
<td>0.3776</td>
<td>0.5404</td>
<td>0.3236</td>
</tr>
<tr>
<td>HBAYES-CS</td>
<td>0.4738</td>
<td>0.5654</td>
<td>0.4639</td>
</tr>
<tr>
<td>KF:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAT</td>
<td>0.2052</td>
<td>0.1293</td>
<td>0.7026</td>
</tr>
<tr>
<td>HTD</td>
<td>0.5800</td>
<td>0.7051</td>
<td>0.5560</td>
</tr>
<tr>
<td>HTD-CS</td>
<td>0.6091</td>
<td>0.6745</td>
<td>0.6156</td>
</tr>
<tr>
<td>HBAYES</td>
<td>0.5512</td>
<td>0.6915</td>
<td>0.5086</td>
</tr>
<tr>
<td>HBAYES-CS</td>
<td>0.6073</td>
<td>0.6759</td>
<td>0.6126</td>
</tr>
</tbody>
</table>

- Best F-score: joint hierarchical cost-sensitive and data fusion techniques
- Best precision: HTD and HBAYES but also HBAYES-CS and HTD-CS perform well
- Best recall: FLAT, but also HBAYES-CS and HTD-CS good results
- Better compromise between precision and recall: HBAYES-CS and HTD-CS.
Conclusions

- Hierarchical strategies show better results than “flat” approaches
- HBAYES-CS and HTD-CS achieve significantly better hierarchical F-scores than the basic HBAYES and HTD ensembles
- This is the result of a better compromise between precision and recall
- With a single global parameter we may tune the precision/recall characteristics of the overall HBAYES-CS ensembles
- Data fusion significantly improve predictions
- We need a synergy between hierarchical, data fusion and cost-sensitive approaches to achieve the best results.
Some open problems ...

- Biomolecular data integration can improve gene function prediction performances: which other methods could be considered?

- Can we extend HBAYES-CS to DAG-structured taxonomies (e.g. GO)?

- Experimental work: comparison with other promising hierarchical ensemble approaches and state-of-the-art methods in the context of genome-wide gene function prediction

- Can these methods to be applied to genome and ontology-wide of multi-cellular eukaryotes? (e.g. *A.thaliana*, mouse or human)
Large room for further research and for improvement...
References


* All these papers (and others) are available from: http://homes.dsi.unimi.it/~valenti/pub.html