Prediction of human gene - phenotype associations by exploiting the hierarchical structure of the Human Phenotype Ontology

Giorgio Valentini[1], Sebastian Köhler[2],
Matteo Re[1], Marco Notaro[3], and Peter N. Robinson[2,4]

[1] AnacletoLab - DI, Dipartimento di Informatica,
Università degli Studi di Milano, Italy {valentini,re}@di.unimi.it
[2] Institut fur Medizinische Genetik und Humangenetik,
Charité - Universitatsmedizin Berlin, Germany
{peter.robinson,sebastian.koehler}@charite.de
[3] Dipartimento di Bioscienze,
Università degli Studi di Milano, Italy marco.notaro@studenti.unimi.it
[4] Institute of Bioinformatics, Department of Mathematics and Computer Science,
Freie Universitat Berlin, Germany

Abstract. The Human Phenotype Ontology (HPO) provides a conceptualization of phenotype information and a tool for the computational analysis of human diseases. It covers a wide range of phenotypic abnormalities encountered in human diseases and its terms (classes) are structured according to a directed acyclic graph. In this context the prediction of the phenotypic abnormalities associated to human genes is a key tool to stratify patients into disease subclasses that share a common biological or pathophisiological basis. Methods are being developed to predict the HPO terms that are associated for a given disease or disease gene, but most such methods adopt a simple "flat" approach, that is they do not take into account the hierarchical relationships of the HPO, thus loosing important a priori information about HPO terms. In this contribution we propose a novel Hierarchical Top-Down (HTD) algorithm that associates a specific learner to each HPO term and then corrects the predictions according to the hierarchical structure of the underlying DAG. Genome-wide experimental results relative to a complex HPO DAG including more than 4000 HPO terms show that the proposed hierarchical-aware approach significantly improves predictions obtained with flat methods, especially in terms of precision/recall results.

Keywords: Human Phenotype Ontology term prediction, Ensemble methods, Hierarchical classification methods, Disease gene prioritization

1 Introduction

The characterization of human diseases through detailed phenotypic data and the ever increasing amount of genomic data available through high-throughput

technologies can improve our understanding of the bio-molecular mechanisms underlying human diseases. Indeed phenotypic analysis is fundamental for our understanding of the pathophysiology of cellular networks and plays a central role in the mapping of disease genes [1].

To this end the Human Phenotype Ontology (HPO) project [2] provides a comprehensive and well-structured set of more than 10000 terms (classes) that represent human phenotypic abnormalities annotated to more than 7000 hereditary syndromes listed in OMIM, Orphanet and DECIPHER databases [3]. This resource offers an ontology, that is, a conceptualization of the human phenotypes that can be processed by computational methods, and provides a translational bridge from genome-scale biology to a disease-centered view of human pathobiology [4]. The HPO provides also hierarchical relationships between terms, representing the *is_a* relation between them, whereby each term may have more than one parent, thus resulting in a Directed-Acyclic-Graph (DAG) structure of the overall ontology.

In this context, the prediction or ranking of genes with respect to HPO terms is an important computational task . This task is related but different from the classical disease-gene prioritization problem, in which genes are prioritized with respect to specific diseases [5]. Indeed we rank genes with respect to HPO terms. Note that HPO terms do not themselves represent diseases, but rather they denote the individual signs and symptoms and other clinical abnormalities that characterize diseases. Thus, one disease is characterized by ≥ 1 HPO term, and many HPO terms are associated with multiple distinct diseases.

Several computational methods have been applied to predict gene - phenotype associations [6, 7, 8, 9], but they do not take into account the hierarchical relationships that characterize phenotypes both in human and model organisms. The resulting "flat" predictions, i.e. predictions unaware of the relationships between the different phenotypes, may provide inconsistent results. For instance, if we adopt the HPO to catalogue human phenotypes and we try to predict HPO terms independently of each other, we could associate to some human gene the HPO term "Atrial septal defect" but not the term "Abnormality of the cardiac septa", thus introducing an inconsistency since "Atrial septal defect" is obviously a subclass of "Abnormality of the cardiac septa". Besides inconsistency, flat predictions loose the available "a priori" knowledge about the hierarchical relationships between HPO terms, thus suggesting that hierarchy-aware methods could at least in principle introduce improvements in the gene-phenotype predictions. To overcome the limitations of "flat" approaches, we could apply computational methods for hierarchically structured output spaces, but most of them have focused on tree-structured ontologies [10, 11, 12, 13] and only a few on DAG-structured taxonomies [14, 15] and, even if they have been applied in computational biology, e.g. to the prediction of protein functions [16], to our knowledge no hierarchy-aware methods have been applied to the prediction of HPO terms associated to human genes.

To fill this gap, we propose a simple and novel hierarchical method, i.e. the *Hierarchical Top-Down (HTD)* ensemble method conceived to deal with the

DAG structure of the HPO. At first a base learner associated with each considered HPO term is applied to provide "flat" gene-phenotype associations. Then the algorithm gradually visits the HPO DAG level by level from the root (top) to the leaves (bottom), and modifies the flat predictions to assure their hierarchical consistency. One of the main advantages of the proposed approach is that it always provides consistent predictions, that is predictions that respect the hierarchical structure of the HPO. Moreover, by exploiting the parent-child relationships between HPO terms, the proposed hierarchical approach can significantly improve HPO flat predictions, as shown by the large set of experiments involving more than $20,000$ human genes and more than $4,000$ HPO terms. The *HTD* method is simple, fast and can be applied by using in principle any base learner for both hierarchical multi-label phenotypic classification and ranking of human disease genes.

2 Hierarchical Top-Down (HTD) ensembles for the HPO taxonomy

Let $G =< V, E >$ be a Directed Acyclic Graph (DAG) with vertices $V = \{1, 2, \ldots, |V|\}$ and edges $e = (i, j) \in E, i, j \in V$. G represents a taxonomy structured as a DAG, whose nodes $i \in V$ represent classes of the taxonomy and a directed edge $(i, j) \in E$ the hierarchical relationships between i and j: i is the parent class and j is the child class. In our experimental setting the unique root node $root(G)$ is represented by the top HPO term "HP:0000001": all the other HPO terms are its descendants. The set of children of a node i is denoted $child(i)$, and the set of its parents $par(i)$.

To each HPO term i is associated a "flat" classifier $f_i : X \rightarrow [0, 1]$ that provides a score $\hat{y}_i \in [0, 1]$ for a given gene $x \in X$. Ideally $\hat{y}_i = 1$ if gene x is associated to the HPO term i, and $\hat{y}_i = 0$ if it is not, but intermediate scores are allowed. The ensemble of the $|V|$ flat classifiers provides a score for each node/class $i \in V$ of the DAG G:

$$\hat{\boldsymbol{y}} =< \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|V|} > \tag{1}$$

We say that the multi-label scoring $\boldsymbol{y}$ is valid if it obeys the *true path rule* (also called the *annotation propagation rule*) that holds also for other DAG-structured ontologies, such as the Gene Ontology (when restricted to subclass relations) [17]:

$$\boldsymbol{y} \text{ is valid } \iff \forall i \in V, j \in par(i) \Rightarrow y_j \geq y_i \tag{2}$$

According to this rule, if we assign a HPO term i to a gene, then also its parent HPO terms must be assigned to the same gene: in other words an assignment to a node must be recursively extended to all its ancestors. Note that this implies that a score for a parent HPO term must be larger or equal than that of its children. Consequently, if a certain HPO term is classified as a negative example

because its score is below threshold, then all of its descendents must also be classified negative.

In real cases it is very unlikely that a flat classifier satisfies the true path rule, since by definition the predictions are performed without considering the hierarchy of the classes. Nevertheless by adding a further label/score modification step, i.e. by taking into account the hierarchy of the classes, we can modify the labeling or the scores of the flat classifiers to obtain a hierarchical classifier that obeys the true path rule.

To this end we propose a Hierarchical top-down algorithm *(HTD)*, that modifies the flat scores according to the hierarchy of a DAG through a unique run across the nodes of the graph. It adopts this simple rule by per-level visiting the nodes from top to bottom:

$$\bar{y}_i := \begin{cases} \hat{y}_i & \text{if} \quad i \in root(G) \\ \min_{j \in par(i)} \bar{y}_j & \text{if} \quad \min_{j \in par(i)} \bar{y}_j < \hat{y}_i \\ \hat{y}_i & \text{otherwise} \end{cases} \tag{3}$$

Note that $\bar{\boldsymbol{y}} =< \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{|V|} >$ represents the set of the predictions obtained by the *(HTD)* algorithm from the flat predictions $\hat{\boldsymbol{y}} =< \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|V|} >$.

The node levels correspond to their maximum path length from the root. More precisely, having $\mathcal{L} = \{0, 1, \ldots, \xi\}$ levels in the HPO taxonomy, $\psi : V \longrightarrow \mathcal{L}$ is a level function which assigns to each HPO term $i \in V$ a level, i.e. its maximum distance from the root. For instance, nodes $\{i|\psi(i) = 0\}$ correspond to the root node, $\{i|\psi(i) = 1\}$ is the set of nodes with a maximum path length from the root (distance) equal to 1, and $\{i|\psi(i) = \xi\}$ are nodes that lie at a maximum distance ξ from the root.

Fig 1 shows that we need to visit the HPO hierarchy per level in the sense of the maximum and not of the minimum distance from the root: this is necessary to preserve the consistency of the predictions. Indeed looking at the *HTD* scores obtained respectively with minimum and maximum distance from the root (bottom-left of Fig. 1), we see that only the maximum distance preserves the consistency of the predictions. Indeed, focusing on node 5, by traversing the DAG levels according to the minimum distance from the root, we have that the level of node 5 is 1 ($\psi^{min}(5) = 1$) and in this case by applying the *HTD* rule (3) the flat score $\hat{y}_5 = 0.8$ is wrongly modified to the *HTD* ensemble score $\bar{y}_5 = 0.7$. If we instead traverse the DAG levels according to the maximum distance from the root, we have $\psi^{max}(5) = 3$ and the *HTD* ensemble score is correctly set to $\bar{y}_5 = 0.4$. In other words at the end of the *HTD*, by traversing the levels according to the minimum distance we have $\bar{y}_5 = 0.7 > \bar{y}_4 = 0.4$, that is a child node has a score larger than that of its parent, and the true path rule is not preserved; on the contrary by traversing the levels according to the maximum distance we achieve $\bar{y}_5 = 0.4 \leq \bar{y}_4 = 0.4$ and the true path rule consistency is assured. This is due to the fact that by adopting the minimum distance when we visit node 5, node 4 has not just been visited, and hence the value 0.4 has not been transmitted by node 2 to node 4; on the contrary if we visit the DAG according to the maximum distance all the ancestors of node 5 (including node

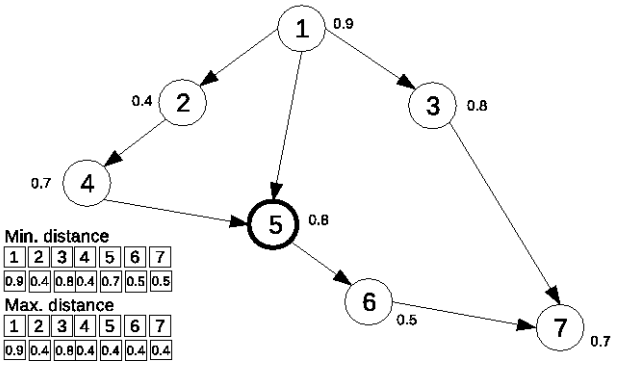

Fig. 1. Levels of the hierarchy must be defined in terms of the maximum distance from the root (node 1). Small numbers close to nodes correspond to the scores of the flat predictions. The Hierarchical top-down scores obtained respectively by crossing the levels according to the minimum and the maximum distance from the root are shown in the bottom-left.

4) have just been visited and the score 0.4 is correctly transmitted to node 5 along the path $2 \rightarrow 4 \rightarrow 5$.

More precisely, given a DAG $G =< V, E >$, the level function ψ, a set of flat predictions $\hat{\boldsymbol{y}} =< \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|V|} >$ for each class associated to each node $i \in \{1, \ldots, |V|\}$, the *HTD-DAG* algorithm assures that for the set of ensemble predictions $\bar{\boldsymbol{y}} =< \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{|V|} >$ the following property holds:

$$\forall i \in V,\ j \in par(i) \Rightarrow \bar{y}_j \geq \bar{y}_i \tag{4}$$

Indeed, by applying the rule (3) from the top to the bottom of the hierarchy we assure that the scores of the parents are larger or equal than those of its children. Moreover by visiting "per level" the hierarchy according to the ψ function (levels are defined in the sense of the maximum distance) we assure that each parent has just been visited before their children and by observing that each node is visited only once it cannot be changed by the successive top-down steps of the algorithm, thus assuring that $\forall i \in V,\ j \in par(i) \Rightarrow \bar{y}_j \geq \bar{y}_i$.

There are several ways to implement the function ψ that computes the maximum distance of each node from the root. We applied the classical Bellman-Ford algorithm [18]: by recalling that it finds the shortest paths from a source node to all the other nodes of a weighted digraph, it is sufficient to invert the sign of each edge weight to obtain the maximum distance (longest path) from the root. We outline that other methods (e.g. procedures based on the topological sort of graphs) are more efficient, but considering that the levels should be computed only once, on modern computers there are not significant differences in terms of the mean empirical computational time.

Fig. 2. Hierarchical Top-Down algorithm for DAGs (HTD)

```
Input:
- G =< V, E >
- ŷ =< ŷ_1, ŷ_2, ..., ŷ_|V| >,   ŷ_i ∈ [0, 1]
begin algorithm
01:    A. Compute ∀i ∈ V the max distance from root(G):
02:        E' := {e'|e ∈ E, e' = −e}
03:        G' :=< V, E' >
04:        dist := Bellman.Ford(G', root(G'))
05:    B. Per-level top-down visit of G:
06:        ȳ_root(G) := ŷ_root(G)
07:        for each d from 1 to max(dist) do
08:          N_d := {i|dist(i) = d}
09:          for each i ∈ N_d do
10:            x := min_{j∈par(i)} ȳ_j
11:            if (x < ŷ_i)
12:              ȳ_i := x
13:            else
14:              ȳ_i := ŷ_i
15:          end for
16:        end for
end algorithm
Output:
- ȳ =< ȳ_1, ȳ_2, ..., ȳ_|V| >
```

Fig. 2 shows the pseudo code of the overall *HTD* algorithm. Rows $1-4$ provide the maximum distance of each node from the root, whereas the block B of the algorithm implements a per-level top-down visit of the graph (rows $5-16$).

Starting from the children of the root (level 1) for each level of the graph the nodes are processed and the hierarchical top-down correction of the flat predictions $\hat{y}_i$, $i \in \{1, \ldots, |V|\}$ to the HTD-DAG ensemble prediction $\bar{y}_i$ is performed according to eq. 3. It is easy to see that the complexity of block B (rows $5-16$) is linear in the number of vertices for sparse graphs (and the HPO is just a sparse DAG).

3 Experimental set-up

3.1 The Human Phenotype Ontology

Ontologies are high-level representations of knowledge domains based upon controlled vocabularies. The Human Phenotype Ontology (HPO) aims at providing a standardized categorization of the abnormalities associated to human diseases (each represented by an HPO term) and the semantic relationships between

them. Each HPO term describes a phenotypic abnormality and is developed using medical literature and cross-references to other biomedical ontologies (e.g. OMIM [3]).

A key feature of HPO is its ability, based upon an equivalence mapping to other publicly available phenotype vocabularies, to allow the integration of existing datasets and to strongly promote the interoperability with multiple biomedical resources [4].

The experiments presented in this manuscript are based on the September 2013 HPO release (10,099 terms and 13,382 between-term relationships). The annotations of the 20,257 human genes were taken from the same HPO release. After pruning the HPO terms having less than 2 annotations we obtained a final HPO DAG composed by 4,847 terms (and 5,925 between-terms relationships)

3.2 Construction and integration of the protein functional network

The set of human genes considered in the experiments presented here was obtained from the recent critical assessment of protein function annotation (CAFA2) international challenge. Starting from an initial set of 20,257 human genes we constructed, for each gene, different binary profile vectors representing the absence/presence of bio-molecular features in the gene product encoded by the considered gene. More precisely, we constructed for each gene 8 binary vectors containing the features obtained, respectively, from InterPro [19], Pfam [20], PRINTS [21], PROSITE [22], SMART [23], SUPFAM [24], Gene Ontology [17] and OMIM [3]. All these annotations were obtained by parsing the raw text annotation files made available by the Uniprot knowledgebase (release May 2013, considering only its SWISSprot component database). We then obtained a similarity score between each pair of genes simply by computing the Jaccard similarity between the feature vectors associated with the genes.

Following this strategy we obtained 8 gene networks (one for each of the aforementioned data sources). The final functional interaction network used in the presented experiments was constructed using a simple unweighted integration strategy that does not involve any learning phase in the network integration process: the *Unweighted Average (UA)* network integration method[25]. In UA the weight of each edge of the combined networks is computed simply averaging across the available n networks:

$$\bar{w}_{ij} = \frac{1}{n} \sum_{d=1}^{n} w_{ij}^{d} \tag{5}$$

In order to construct a more informative network, we added also two more functional gene networks taken from the literature and previously published in [26, 27].

3.3 Kernelized score functions

As base learner we used a semi-supervised network-based learning method recently successfully applied to gene disease prioritization [28], gene function pre-

diction [29] and drug repositioning [30]. Kernelized score functions adopt both a local and a global learning strategy. Local learning is accomplished through a generalization of the classical guilt-by-association approach [31], through the introduction of different functions to quantify the similarity between a gene and its neighbours. A global learning strategy is introduced in form of a kernel that can capture the overall topology of the underlying biomolecular network.

More precisely, by this approach we can derive score functions $S : V \longrightarrow \mathbb{R}^+$ based on properly chosen kernel functions, by which we can directly rank a gene v according to the values of $S(v)$: the higher the score, the higher the likelihood that a gene belongs to a given class [29]. The score functions are built on distance measures defined in a suitable Hilbert space $\mathcal{H}$ and computed using the usual "kernel trick", by which instead of explicitly computing the inner product $< \phi(\cdot), \phi(\cdot) >$ in the Hilbert space, with $\phi : V \longrightarrow \mathcal{H}$, we compute the associated kernel function $K : V \times V \longrightarrow \mathbb{R}^+$ in the original input space V.

For instance, given a vertex v, a set of genes V_C belonging to a specific class C, we can obtain the following *Average score* S_{AV}:

$$S_{AV}(v, V_C) = \frac{1}{|V_C|} \sum_{x \in V_C} K(v, x) \tag{6}$$

In principle any valid kernel K can be applied to compute the aforementioned kernelized score, but in the context of gene - phenotype association ranking, we used *random walk kernels* [32], since they can capture the similarity between genes, taking into account the topology of the overall functional interaction network.

In our experiments we applied a *1-step random walk kernel*: in this way we explicitly evaluate only the direct neighbors of each gene in the functional interaction network. It is worth noting that other kernels may lead to better results, but here we are mainly interested in verifying whether our proposed *HTD* algorithm can improve upon Flat predictions, and not in fine tuning and achieving the best possible results.

4 Results

We compared our proposed *HTD* ensemble methods with flat predictions obtained with 1-step random walk kernelized score functions, by applying classical leave-one-out techniques.

In terms of the average AUC across the 4846 considered HPO terms, even if the difference in favour of *HTD* is very small (0.7923 vs 0.7897), by looking at the results of the single HPO terms, *HTD* improves over flat in 3346 HPO terms, achieves the same AUC for 554 HPO terms and "looses" in 956 terms. This means that for more than 3/4 HPO terms we obtain an improvement, and this explains also why, according to the Wilcoxon rank sum test the difference between the methods is statistically significant in favour of *HTD* at 10^{-5} significance level.

Also better results are obtained when we consider the precision at a fixed recall level. Indeed in this case the average values across HPO terms are quite

Table 1. Average AUC, and precision at 10, 20 and 40% recall (P10R, P20R and P40R). Flat stands for flat ensemble method, HTD for Hierarchical Top-Down, Max for Hierarchical Maximum, And for Hierarchical And and Or for Hierarchical Or ensemble methods. Methods that are significantly better than all the others according to the Wilcoxon rank sum test ($\alpha = 10^{-5}$) are highlighted in bold.

	Flat	**HTD**	**Max**	**And**	**Or**
AUC	0.7897	0.7923	0.7879	**0.8151**	0.7880
P10R	0.1620	**0.1957**	0.1315	0.1665	0.1352
P20R	0.1278	**0.1535**	0.1081	0.1283	0.1110
P40R	0.0812	**0.0890**	0.0728	0.0758	0.0741

consistent: for instance the average precision at 20% recall is 0.1535 vs 0.1278, and another time for most HPO terms we obtain a significant increment when the *HTD* hierarchical correction is applied to the flat predictions.

Table 1 summarizes the average results across terms for the *HTD*, *Flat* and three heuristic hierarchical ensemble methods originally proposed for the hierarchical prediction of Gene Ontology terms [14]. *HTD* achieves always significantly better results than the flat approach, both in terms of AUC and precision at a fixed recall. Moreover it obtains significantly better results than all the other compared hierarchical ensemble methods. The only exception is with respect to the AUC where the *And* hierarchical method achieves better results.

Fig. 3 compares the precision at different recall levels for all the hierarchical and the flat ensemble methods: the *HTD* solid line marked with circles is consistently above all the other curves, showing that *HTD* achieves on the average better results than all the other competing methods.

Even if the average precision at a fixed recall rate is relatively low with all the methods we presented (Fig. 3), we note that we tried to perform predictions also with terms having only two positive annotations, a very difficult task that likely leads in most case to precision values very close to 0. Moreover by applying score functions with 2-step or more random walk kernels, we could better exploit the overall topology of the network and at least potentially achieve better results, especially with terms having a small number of annotations or with "positive" nodes relatively "far" from each other. In any case, at least for low values of recall, *HTD* shows on the average relative improvements of the precision between 10 and 20%, with respect to the *Flat* approach. On the contrary, the heuristic *Max*, *And* and *Or* methods are not able to outperform the *Flat* approach, confirming previous results in the context of gene function prediction [14].

5 Conclusions

The prediction of human gene–abnormal phenotype associations is an important step toward the discovery of novel disease genes associated with hereditary disorders. Several computational methods that exploit "omics" data can be successfully applied to predict or rank genes with respect to human phenotypes, but

usually their predictions are inconsistent, in the sense that do not necessarily obey the parent-child relationships between HPO terms (i.e. a gene may achieve a score for a child term larger than that that of its parent HPO term).

We showed that our proposed method provides predictions that are always consistent, according the "true path rule" that governs the HPO taxonomy. Moreover the *HTD* ensemble method can enhance "flat" predictions by exploiting the hierarchical relationships between HPO terms. Indeed our experimental results showed that *HTD*, by using kernelized score functions as base learner, can significantly improve the precision-recall curves. We obtained a significant increment using also other base learners (e.g. the classical label propagation algorithm described in [33] – data not shown), and in principle our proposed hierarchical method is independent of the base learner used to provide the initial "flat" scores. From this standpoint *HTD* can be applied to improve the performance of any "flat" learning method, and to provide consistent and more reliable predictions for novel gene - phenotype predictions by exploiting the DAG structure of the Human Phenotype Ontology.

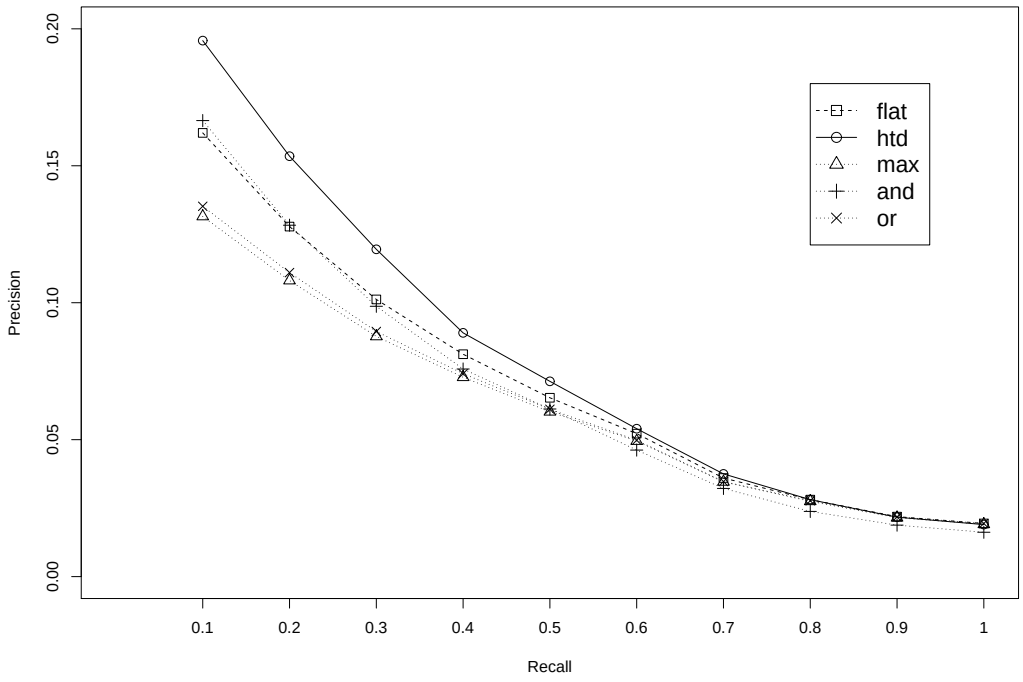

Fig. 3. Precision at different levels of recall, averaged across HPO terms.

Acknowledgments

G.V. and M.R. acknowledge partial support from the PRIN project "Automi e linguaggi formali: aspetti matematici e applicativi", funded by the Italian Ministry of University.

References

[1] Robinson, P., Krawitz, P., Mundlos, S.: Strategies for exome and genome sequence data analysis in disease-gene discovery projects. Cin. Genet. **80** (2011) 127 – 132
[2] Robinson, P., Kohler, S., Bauer, S., Seelow, D., Horn, D., Mundlos, S.: The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am. J. Hum. Genet. **83** (2008) 610–615
[3] Amberger, J., Bocchini, C., Amosh, A.: A new face and new challenges for Online Mendelian inheritance in Man (OMIM). Hum. Mutat. **32** (2011) 564–7
[4] Kohler, S., et al.: The human phenotype ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Research **42**((Database issue)) (2014) D966–74
[5] Moreau, Y., Tranchevent, L.: Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Rev. Genet. **13**(8) (2012) 523–536
[6] McGary, K., Lee, I., Marcotte, E.: Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes. Genome Biology **8**(R258) (2007)
[7] Mehan, M., Nunez-Iglesias, J., Dai, C., Waterman, M., Zhou, X.: An integrative modular approach to systematically predict gene-phenotype associations. BMC Bioinformatics **11**(Suppl 1) (2010)
[8] Wang, P., et al.: Inference of gene-phenotype associations via protein-protein interaction and orthology. PLoS ONE **8**(10) (2013)
[9] Musso, G., et al.: Novel cardiovascular gene functions revealed via systematic phenotype prediction in zebrafish. Development **141** (2014) 224–235
[10] Cerri, R., de Carvalho, A.: Hierarchical multilabel protein function prediction using local neural networks. In: Advances in Bioinformatics and Computational Biology. Number 6832 in Lecture Notes in Computer Science. Springer-Verlag (2011) 10–17
[11] Silla, C., Freitas, A.: A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery **22**(1-2) (2011) 31–72
[12] Valentini, G.: True Path Rule hierarchical ensembles for genome-wide gene function prediction. IEEE ACM Transactions on Computational Biology and Bioinformatics **8**(3) (2011) 832–847
[13] Cesa-Bianchi, N., Re, M., Valentini, G.: Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference. Machine Learning **88**(1) (2012) 209–241
[14] Obozinski, G., Lanckriet, G., Grant, C., M., J., Noble, W.: Consistent probabilistic output for protein function prediction. Genome Biology **9**(S6) (2008)
[15] Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Dzeroski, S.: Predicting gene function using hierarchical multi-label decision tree ensembles. BMC Bioinformatics **11**(2) (2010)
[16] Valentini, G.: Hierarchical Ensemble Methods for Protein Function Prediction. ISRN Bioinformatics **2014**(Article ID 901419) (2014) 34 pages

[17] Gene Ontology Consortium: Gene Ontology annotations and resources. Nucleic Acids Research **41** (2013) D530–535
[18] Cormen, T., Leiserson, C., Rivest, R.: Introduction to Algorithms. MIT Press, Boston (2009)
[19] Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., et al.: The interpro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Research **29**(1) (2001) 37–40
[20] Finn, R., Tate, J., Mistry, J., Coggill, P., Sammut, J., Hotz, H., Ceric, G., Forslund, K., Eddy, S., Sonnhammer, E., Bateman, A.: The Pfam protein families database. Nucleic Acids Research **36** (2008) D281–D288
[21] Attwood, T.: The prints database: a resource for identification of protein families. Brief Bioinform. **3**(3) (2002) 252–263
[22] Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B., De Castro, E., Lachaize, C., Langendijk-Genevaux, P., Sigrist, C.: The 20 years of prosite. Nucleic Acids Research **36** (2008) D245–D249
[23] Schultz, J., Milpetz, F., Bork, P., Ponting, C.: Smart, a simple modular architecture research tool: identification of signaling domains. Proceedings of the National Academy of Sciences **95**(11) (1998) 5857–5864
[24] Gough, J., Karplus, K., Hughey, R., Chothia, C.: Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. Journal of molecular biology **313**(4) (2001) 903–919
[25] Valentini, G., Paccanaro, A., Caniza, H., Romero, A., Re, M.: An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. Artificial Intelligence in Medicine **61**(2) (2014) 63–78
[26] Wu, G., Feng, X., Stein, L.: A human functional protein interaction network and its application to cancer data analysis. Genome Biol **11** (2010) R53
[27] Lee, I., Blom, U., Wang, P.I., Shim, J., Marcotte, E.: Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Research **21**(7) (2011) 1109–1121
[28] Re, M., Valentini, G.: Cancer module genes ranking using kernelized score functions. BMC Bioinformatics **13**(Suppl 14/S3) (2012)
[29] Re, M., Mesiti, M., Valentini, G.: A Fast Ranking Algorithm for Predicting Gene Functions in Biomolecular Networks. IEEE ACM Transactions on Computational Biology and Bioinformatics **9**(6) (2012) 1812–1818
[30] Re, M., Valentini, G.: Network-based Drug Ranking and Repositioning with respect to DrugBank Therapeutic Categories. IEEE/ACM Transactions on Computational Biology and Bioinformatics **10**(6) (2013) 1359–1371
[31] Oliver, S.: Guilt-by-association goes global. Nature **403** (2000) 601–603
[32] Smola, A., Kondor, I.: Kernel and regularization on graphs. In Scholkopf, B., Warmuth, M., eds.: Proc. of the Annual Conf. on Computational Learning Theory. Lecture Notes in Computer Science, Springer (2003) 144–158
[33] Zhu, X., et al.: Semi-supervised learning with gaussian fields and harmonic functions. In: Proc. of the 20th Int. Conf. on Machine Learning, Washintgton DC, USA (2003)