Ensembling Descendant Term Classifiers to Improve Gene – Abnormal Phenotype Predictions

Marco Notaro[(1)], Max Schubach[(2)], Marco Frasca[(1)], Marco Mesiti[(1)], Peter N. Robinson[(3)] and Giorgio Valentini[(1)]

(1) Anacleto Lab – Dipartimento di Informatica, Universitá degli Studi di Milano, Via Comelico 39, 20135, marco.notaro@unimi.it, giorgio.valentini@unimi.it

(2) Institute for Medical and Human Genetics, Charité - Universitätsmedizin Berlin, Augustenburger Platz 1, 13353, Berlin, Germany, max.schubach@charite.de

(3) The Jackson Laboratory for Genomic Medicine, 10 Discovery Dr, CT 06032, Farmington, USA, Peter.Robinson@jax.org

Keywords: Human Phenotype Ontology, Hierarchical Multi-Label Classification, Hierarchical Ensemble Methods, Gene-Abnormal Phenotype Association.

Abstract. The Human Phenotype Ontology (*HPO*) provides a standard categorization of the phenotypic abnormalities encountered in human diseases and of the semantic relationship between them. Quite surprisingly the problem of the automated prediction of the association between genes and abnormal human phenotypes has been widely overlooked, even if this issue represents an important step toward the characterization of gene-disease associations, especially when no or very limited knowledge is available about the genetic etiology of the disease under study. We present a novel ensemble method able to capture the hierarchical relationships between *HPO* terms, and that we show to improve existing hierarchical ensemble algorithms by explicitly considering the predictions of the descendant terms of the ontology. In this way the algorithm exploits the information embedded in the most specific ontology terms that closely characterize the phenotypic information associated with each human gene. Genome-wide results obtained by integrating multiple sources of information show the effectiveness of the proposed approach.

1 Scientific Background

The Human Phenotype Ontology (*HPO*) project [1] aims at providing a standard categorization of the abnormalities associated with human diseases and the semantic relationships between them. Each *HPO* term does not represent a disease, but rather it denotes individual signs or symptoms or other clinical abnormalities that characterize a disease. The *HPO* is structured as a direct acyclic graph (*DAG*), where more general terms are found on the top levels of hierarchy and the term specificity increases moving from root to leaves. All the *HPO* relationships are *is-a*, i.e. class-subclass relationships. While the problem of the prediction of gene–disease associations has been widely investigated [2], the related problem of gene– phenotypic feature (i.e. *HPO* term) association has been largely overlooked, despite the quickly growing application of the HPO to relevant medical problems [3, 4]. In principle in the contest of gene–abnormal phenotype prediction, any "flat" method, that predicts labels independently of each other, can be applied [5], but they may introduce significant inconsistencies in the classification due to the violation of the *true path rule* that governs the *HPO* [1] taxonomy. Besides inconsistency, flat methods may also lose important a priori knowledge about the constraints of the hierarchical labeling that could enhance the accuracy of the predictions. To overcome these limitations we propose an improved extension of the ensemble algorithm

Hierarchical True path Rule for Directed Acyclic Graph (*TPR-DAG*) [6]. We named this novel variant *DEScendant Classifier ENSemble* (*DESCENS*). The novelty of *DESCENS* with respect to *TPR-DAG* consists in strongly considering the contribution of all the descendants of each node instead of only that of its children, since with the *TPR-DAG* algorithm the contribution of the descendants of a given node decays exponentially with their distance from the node itself, thus reducing the impact of the predictions made at the most specific levels of the ontology [7]. On the contrary *DESCENS* predictions are more influenced by the information embedded in the most specific terms of the taxonomy (e.g. leaf nodes), thus putting more emphasis on the terms that most characterize the gene under study.

2 Materials and Methods

Let $G =< V, E >$ 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 the HPO taxonomy structured as a DAG, whose nodes $i \in V$ represent classes (terms) of the ontology and a directed edge $(i, j) \in E$ the hierarchical relationships between i (parent term) and j (child term). A "continuous flat multi-label scoring" predictor $f : X \rightarrow [0, 1]$ provides a score $\hat{y}_i \in [0, 1]$ that can be interpreted as the likelihood or probability that a given gene belongs to a given node/HPO term $i \in V$ of the DAG G. The set of $|V|$ flat classifiers provides a multi-label score $\hat{\boldsymbol{y}} \in [0, 1]^{|V|}$: $\hat{\boldsymbol{y}} =< \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|V|} >$. We say that a multi-label scoring $\boldsymbol{y}$ is consistent if it obeys the *true path rule*:

$$\boldsymbol{y} \text{ is consistent } \iff \forall i \in V, j \in parents(i) \Rightarrow y_j \geq y_i \quad (1)$$

According to this rule the score of a parent or an ancestor node must be larger or equal than that of its children or descendants nodes.

To process and provide flat scores of the proposed hierarchical ensemble methods we used both a semi-supervised network-based approach (*RANKS* [8]) and a supervised machine learning method (Support Vector Machine – *SVM*). In our experiments we applied *RANKS* with the *average score function* and the *random walk kernel* at 1, 2 and 3 steps, i.e. kernels able to evaluate the direct neighbors and those far away 2 and 3 steps from each gene in the network. It is worth noting that *RANKS* returns a score and not a probability [9]. To make the scores comparable across classes we normalized the scores in the sense of the maximum (i.e. we divided the score values of each class by the maximum score of that class) or according to the quantile normalization [10].

After the learning phase the "flat" predictions are modified by the *DESCENS* algorithm, whose high-level pseudo-code is shown in Fig. 1. The block A of the algorithm (row 1) computes the maximum distance of each node from the root. To this end a method based on the Topological Sorting algorithm can be applied [11]. The block B computes a per-level bottom-up visit of the graph G (rows $2 - 9$) to propagate the "positive" predictions across the hierarchy. More precisely, according to the true path rule, only the "positive" descendants of a certain node i (e.g. descendant nodes having scores larger than that of their ancestor node i) influence the prediction for the node i itself (row 6 of Fig. 1). In this way all the "positive" descendants of node i provide the same contribution to the ensemble prediction $\bar{y}_i$. In this context a key issue is the selection of the "positive" descendants Δ_i and to this end different strategies can be applied:

1. *Threshold Free (TF) Strategy.* We choose as "positive" descendants those nodes that achieve a score higher than that of their ancestor node i:

$$\Delta_i := \{j \in desc(i) | \bar{y}_j > \hat{y}_i\} \quad (2)$$

   This strategy leads to the *DESCENS*-TF algorithm (Fig. 1).

Figure 1: DEScendant Classifier ENSemble for DAGs (DESCENS)

```
Input:
- $G =< V, E >$
- $V = \{1, 2, \ldots, |V|\}$
- $\hat{\boldsymbol{y}} =< \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|V|} >, \quad \hat{y}_i \in [0, 1]$
begin algorithm
01:  A. $dist := \forall i \in V$ ComputeMaxDist $(G, root(G))$
02:  B. Per-level bottom-up visit of $G$:
03:     for each $d$ from $\max(dist)$ to 0 do
04:       $N_d := \{i | dist(i) = d\}$
05:       for each $i \in N_d$ do
06:         $\Delta_i := \{j \in desc(i) | \bar{y}_j > \hat{y}_i\}$
07:         $\bar{y}_i := \frac{1}{1+|\Delta_i|}(\hat{y}_i + \sum_{j \in \Delta_i} \bar{y}_j)$
08:       end for
09:     end for
10:  C. Per-level top-down visit of $G$:
11:     $\hat{\boldsymbol{y}} := \bar{\boldsymbol{y}}$
12:     for each $d$ from 1 to $\max(dist)$ do
13:       $N_d := \{i | dist(i) = d\}$
14:       for each $i \in N_d$ do
15:         $x := \min_{j \in parents(i)} \bar{y}_j$
16:         if $(x < \hat{y}_i)$
17:           $\bar{y}_i := x$
18:         else
19:           $\bar{y}_i := \hat{y}_i$
20:       end for
21:     end for
end algorithm
Output:
- $\bar{\boldsymbol{y}} =< \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{|V|} >$
```

2. *Adaptive Threshold (T) Strategy.* The threshold is selected to maximize some performance metric $\mathcal{M}(j, t)$ (e.g. F-score or *AUPRC*) estimated on the training data for the class j with respect to the threshold t. The corresponding set of positives $\forall i \in V$ is:

$$\Delta_i := \{j \in desc(i) | \bar{y}_j > t_j^*, t_j^* = \arg\max_t \mathcal{M}(j, t)\} \tag{3}$$

   For instance t_j^* can be selected from a set of $t \in (0, 1)$ through internal cross-validation techniques. This strategy leads to the *DESCENS*-T algorithm, simply by changing row 6 with eq. 3 in Fig. 1.

Moreover, by changing line 7 of the pseudo-code shown in Fig 1, we can design the "weighted" version of the *DESCENS* algorithm (*DESCENS*-W) merely adding a weight $w \in [0, 1]$ to balance the contribution between the node i and that of its "positive" descendants:

$$\bar{y}_i := w\hat{y}_i + \frac{(1-w)}{|\Delta_i|} \sum_{j \in \Delta_i} \bar{y}_j \tag{4}$$

Another variant of *DESCENS* (named *DESCENS*-τ) balances the contribution between the "positive" children of a node i and that of its "positive" descendants excluding the children by adding a weight $\tau \in [0,1]$:

$$\bar{y}_i := \frac{\tau}{1+|\phi_i|}(\hat{y}_i + \sum_{j\in\phi_i} \bar{y}_j) + \frac{1-\tau}{1+|\delta_i|}(\hat{y}_i + \sum_{j\in\delta_i} \bar{y}_j) \tag{5}$$

where ϕ_i are the "positive" children of i and $\delta_i = \Delta_i \setminus \phi_i$ the descendants of i without its children. If $\tau = 1$ we consider only the contribution of the "positive" children of i, and if $\tau = 0$ only the descendants that are not children contribute to the score, while for intermediate values of τ we can balance the contribution of ϕ_i and δ_i positive nodes.

Independently of which variants of the *DESCENS* algorithm we decide to use, "positive" predictions are recursively "bottom-up" propagated from the parents towards the ancestors of each node. It is worth nothing that the bottom-up step does not assure the consistency of the predictions. This is guarantee by the block C of the algorithm (row $10-21$), where the nodes are top-down processed by level in an increasing order (from the least to the most specific terms) and the "bottom-up" scores computed at the block B are hierarchically corrected to $\bar{y}$ according to the following simple rule:

$$\bar{y}_i := \begin{cases} \hat{y}_i & \text{if} \quad i \in root(G) \\ \min_{j\in parents(i)} \bar{y}_j & \text{if} \quad \min_{j\in parents(i)} \bar{y}_j < \hat{y}_i \\ \hat{y}_i & \text{otherwise} \end{cases} \tag{6}$$

The aim of the top-down step consists in propagating the "negative" predictions towards the children and in a recursive way towards the descendants of each node. Considering the sparseness of the *HPO*, it is easy to see that the overall computational complexity of *DESCENS* algorithm is $\mathcal{O}(|V|)$.

3 Results

We downloaded physical and genetic experimental interactions relative to 4970 proteins from BioGRID (v. 3.2.106, [12]) and the integrated protein-protein interaction and functional association data for 18172 human proteins from STRING (v. 9.1, [13]). Moreover, starting from the Gene Ontology annotations for the three main sub-ontologies (Biological Process, Molecular Function and Cellular Component) and from OMIM annotations [14], both represented as binary feature vectors, we constructed 4 more networks by using the classical Jaccard index to represent the edge weight (functional similarity) between the nodes (genes) of the resulting network. All these annotations were obtained by parsing the raw text annotation files made available by Uniprot knowledgebase considering only its SWISSPROT component. Finally the resulting $n = 6$ networks have been integrated by averaging the edge weights w_{ij}^d between the genes i and j of each network $d \in \{1, n\}$ after normalizing their weights in the same range of values $w_{ij}^d \in [0,1]$ (*Unweighted Average* (UA) network integration, [15]):

$$\bar{w}_{ij} = \frac{1}{n}\sum_{d=1}^{n} w_{ij}^d \tag{7}$$

The resulting weighted adjacency matrix representing the obtained networks is made up of $19,430$ human proteins.

From the *HPO* website we download the January 2014 release, by considering the *Phenotypic Abnormality* subontology, that is the main subontology of the *HPO* (the other subontologies are significantly smaller and amount to only some tens of terms). To avoid prediction of *HPO* terms having too few annotations, for a reliable assessment we pruned *HPO* terms having less than 10 annotations obtaining a final *HPO-DAG* composed by 2154 *HPO* terms and 2641 between-terms-relationship.

The generalization performance of the methods were assessed through a classical 5-fold cross-validation procedure, whereas the results were evaluated by using the *gene-centric* metric F_{max} (i.e. the maximum hierarchical F-score achievable by "a posteriori" setting an optimal decision threshold [16]) and two *term-centric* metrics: the classical Area Under the Receiver Operating Characteristic Curve (*AUROC*) and the Area Under the Precision Recall Curve (*AUPRC*) to take into account the imbalance of annotated vs. unannotated *HPO* terms.

Table 1 summarizes the results achieved by the hierarchical methods *HTD-DAG* [17] and *TPR-DAG* [6] and by *DESCENS*, the novel ensemble variant presented in this manuscript.

Table 1: Average *AUROC* and *AUPRC* across terms and average F_{max}, Precision and Recall across genes of *HTD-DAG*, *TPR-DAG* and *DESCENS* ensemble variants using both *RANKS* and *SVMs* as base learner. Results of "flat" *RANKS* and *SVMs* are also reported. Results are estimated through 5-fold cross-validation. Separately for each metric and base learner the results significantly better than the others according to the Wilcoxon Rank Sum Test ($\alpha = 10^{-9}$) are highlighted in bold.

Method	**AUROC**	**AUPRC**	$\mathbf{F_{max}}$	**Precision**	**Recall**
RANKS (flat)	0.8493	0.0910	0.3106	0.2407	0.4377
HTD-RANKS	0.8506	0.1065	0.3411	0.2717	0.4583
TPR-TF-RANKS	**0.8567**	0.1166	0.3547	0.2880	**0.4615**
TPR-T-RANKS	0.8512	0.1338	0.3574	0.2929	0.4582
TPR-W-RANKS	0.8507	0.1264	0.3620	0.3025	0.4506
DESCENS-TF-RANKS	0.8554	0.1082	0.3679	0.3148	0.4426
DESCENS-τ-RANKS	0.8530	**0.1360**	0.3622	0.3021	0.4520
DESCENS-T-RANKS	0.8503	0.1087	**0.3771**	**0.3227**	0.4535
DESCENS-W-RANKS	0.8502	0.1223	0.3671	0.3071	0.4561
SVM (flat)	0.7128	0.0429	0.1205	0.1165	0.1247
HTD-SVM	**0.8328**	0.0888	0.2597	0.1898	**0.4112**
TPR-TF-SVM	0.7060	0.0525	0.2034	0.1633	0.2694
TPR-T-SVM	0.8297	**0.1036**	0.2611	0.1939	0.3997
TPR-W-SVM	0.7915	0.0909	0.2187	0.1827	0.2723
DESCENS-TF-SVM	0.7092	0.0561	0.2338	0.1877	0.3100
DESCENS-τ-SVM	0.7182	0.0666	0.2424	0.1927	0.3266
DESCENS-T-SVM	0.7940	0.0514	**0.3102**	**0.2796**	0.3483
DESCENS-W-SVM	0.7724	0.0948	0.2373	0.1815	0.3427

It is worth noting that all the hierarchical ensemble methods are always able to improve the results of the flat methods used as base learner both in terms of *AUROC*, *AUPRC* and F_{max}. More in detail looking at the results obtained using *RANKS* as base learner, *DESCENS*-τ and *DESCENS*-T achieve better results than all the other compared methods in terms of *AUPRC* and F_{max}, while *TPR-TF* achieves the best results in terms of *AUROC*, but *HPO* classes are highly imbalanced, and in this setting it is well-known that *AUPRC* is a significantly more reliable metric than *AUROC* [18]. Looking at the results obtained using as base learner the *SVMs*, we can observe that, independently of the ensemble method chosen, we achieve a significant strong improvement respect to the flat prediction, especially in terms of *AUPRC* and F_{max}. Interestingly enough, considering F_{max}, the only hierarchical metric among those considered, *DESCENS* achieves significantly better results both if we use *RANKS* or *SVMs* as base learners. Finally from Table 1 we can observe how the performances of hierarchical ensembles largely depend on that of the flat base learner: for instance *DESCENS*-τ-*RANKS* achieves a significantly larger average *AUPRC* than *DESCENS*-τ-*SVM*. This is not surprising since the improvement introduced by hierarchical ensemble methods also depends on the the predictions of the underlying flat base learner (Fig. 2).

4 Conclusion

Genome and ontology wide experimental results show that the *DESCENS* algorithm is able to improve the predictions of both semi-supervised flat methods, such as the *RANKS* algorithm, that resulted one of the top ranked method in the recent *CAFA2* challenge for *HPO* term prediction [16], and of supervised methods such as *SVMs*, in terms of *AUROC*, *AUPRC* and F_{max}. Moreover *DESCENS* further improves *HTD-DAG*, and *TPR-DAG* in terms of both *AUPRC* and F_{max}. Furthermore the proposed ensemble methods always provide consistent predictions that obey the *true path rule*, a fundamental fact to assure biologically coherent predictions among *HPO* terms.

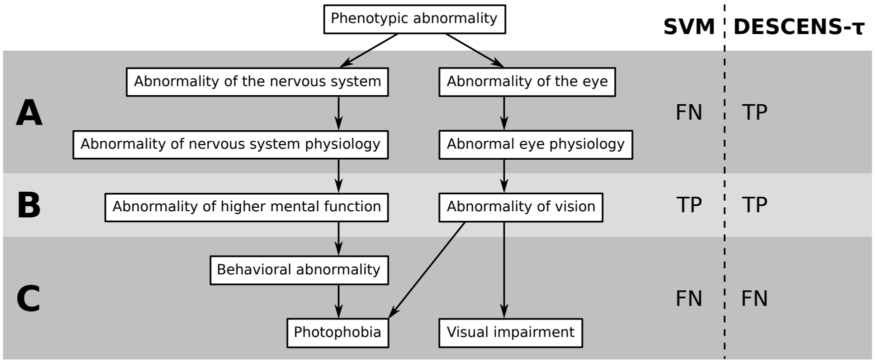

Figure 2: Flat and hierarchical *DESCENS*-τ *HPO* predictions for the gene *RGS9*. At the right side are displayed the correct TP and the incorrect FN predictions made respective by flat-*SVM* and by hierarchical *DESCENS*-τ. In the A box are depicted the predictions that the hierarchical method was able to correct with respect to flat method ($FN \rightarrow TP$); in the B are portrayed the correct predictions for both flat and hierarchical methods and finally in C are shown the incorrect flat predictions that the hierarchical method was not able to recover.

References

[1] N. Kohler, S.and Vasilevsky, M. Engelstad, *et al.*, “The Human Phenotype Ontology in 2017,” *Nucleic Acids Research*, vol. 45, p. D865, 2017.

[2] Y. Moreau and L. Tranchevent, “Computational tools for prioritizing candidate genes: boosting disease gene discovery.,” *Nature Rev. Genet.*, vol. 13, pp. 523–536, 2012.

[3] T. Zemojtel *et al.*, “Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome.,” *Sci Transl Med*, vol. 6, p. 252ra123, 2014.

[4] D. Smedley *et al.*, “A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease.,” *The American Journal of Human Genetics*, vol. 99, pp. 595–606, 2016.

[5] P. Wang *et al.*, “Inference of gene-phenotype associations via protein-protein interaction and orthology.,” *PLoS ONE*, vol. 8, pp. 1–8, 2013.

[6] P. Robinson, M. Frasca, S. Köhler, M. Notaro, M. Re, and G. Valentini, “A hierarchical ensemble method for dag-structured taxonomies,” in *MCS 2015*, Lecture Notes in Computer Science, pp. 15–26, Springer, 2015.

[7] G. Valentini, “True Path Rule hierarchical ensembles for genome-wide gene function prediction,” *IEEE ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, pp. 832–847, 2011.

[8] G. Valentini, G. Armano, M. Frasca, J. Lin, M. Mesiti, and M. Re, “RANKS: a flexible tool for node label ranking and classification in biological networks,” *Bioinformatics*, vol. 32, p. 2872, 2016.

[9] M. Re, M. Mesiti, and G. Valentini, “A fast ranking algorithm for predicting gene functions in biomolecular networks,” *IEEE ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, pp. 1812–1818, 2012.

[10] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias,” *Bioinformatics*, vol. 19, pp. 185–193, 2003.

[11] T. Cormen, C. Leiserson, R. Rivest, and S. RL, *Introduction to Algorithms*. Boston: MIT Press, 2009.

[12] A. Chatr-Aryamontri, B.-J. Breitkreutz, S. Heinicke, L. Boucher, A. G. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. O’Donnell, T. Reguly, A. Breitkreutz, A. Sellam, D. Chen, C. Chang, J. M. Rust, M. S. Livstone, R. Oughtred, K. Dolinski, and M. Tyers, “The BioGRID interaction database: 2013 update.,” *Nucleic Acids Research*, vol. 41, pp. 816–823, 2013.

[13] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, and L. J. Jensen, “STRING v9.1: protein-protein interaction networks, with increased coverage and integration,” *Nucleic Acids Research*, vol. 41, pp. 808–815, 2013.

[14] J. Amberger, C. Bocchini, and A. Amosh, “A new face and new challenges for online mendelian inheritance in man (OMIM),” *Hum. Mutat.*, vol. 32, pp. 564–7, 2011.

[15] G. Valentini, A. Paccanaro, H. Caniza, A. Romero, and M. Re, “An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods,” *Artificial Intelligence in Medicine*, vol. 61, pp. 63–78, 2014.

[16] Y. Jiang *et al.*, “An expanded evaluation of protein function prediction methods shows an improvement in accuracy,” *Genome Biology*, vol. 17, p. 184, 2016.

[17] G. Valentini, S. Köhler, M. Re, M. Notaro, and P. Robinson, “Prediction of human gene - phenotype associations by exploiting the hierarchical structure of the human phenotype ontology,” vol. 9043 of *Lecture Notes in Computer Science*, pp. 66–77, Springer, 2015.

[18] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets,” *PLOS ONE*, vol. 10, pp. 1–21, 03 2015.