Royal Holloway *10th March 2015*

Multi-label hierarchical prediction methods and their application to the automatic function prediction of proteins

Giorgio Valentini

Computer Science Department

UNIVERSITÀ
DEGLI STUDI
DI MILANO

Anacleto Lab

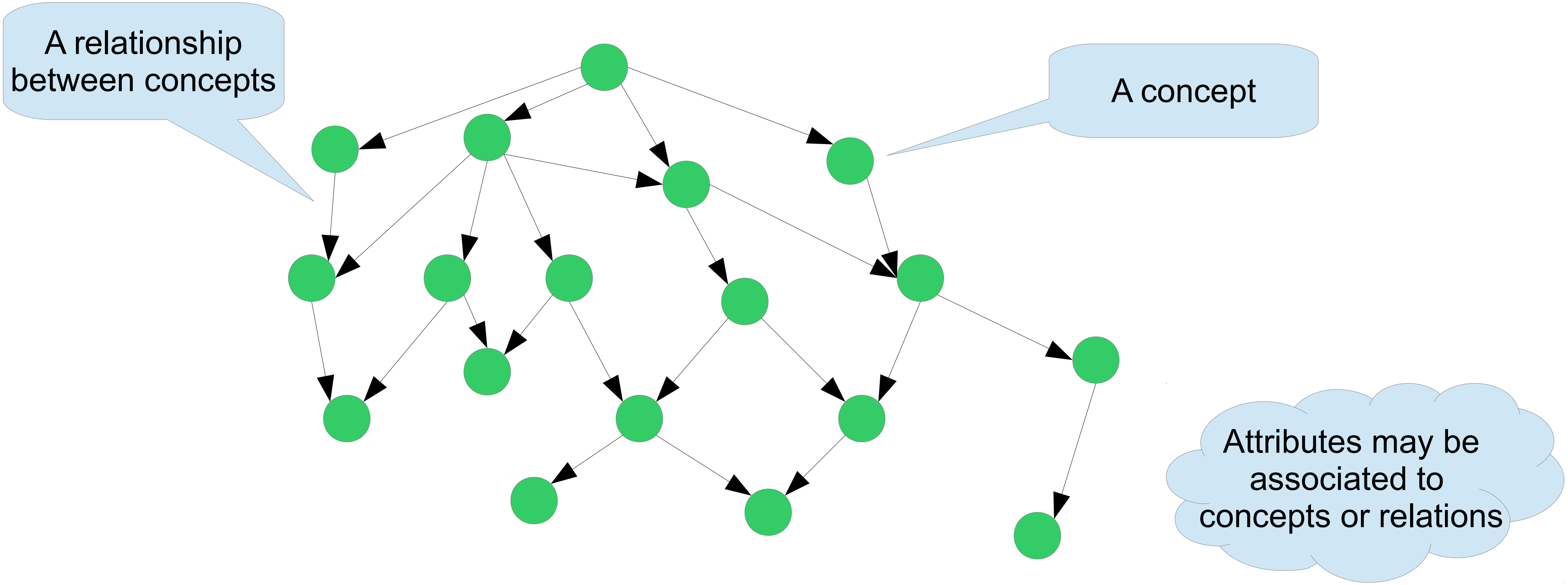
Computational Biology and Bioinformatics

Outline

- Relevant problems in molecular biology and medicine can be modeled through ontologies.
- An example: the Automatic Function Prediction (AFP) problem
- Flat vs Hierarchy-aware learning methods
- Hierarchical ensemble methods
- Perspectives

Ontologies

An ontology is a data model in a given knowledge domain that represents concepts, attributes and relationships in the form of a Directed Acyclic Graph (DAG)



A relationship between concepts

A concept

Attributes may be associated to concepts or relations

Bio-Ontologies

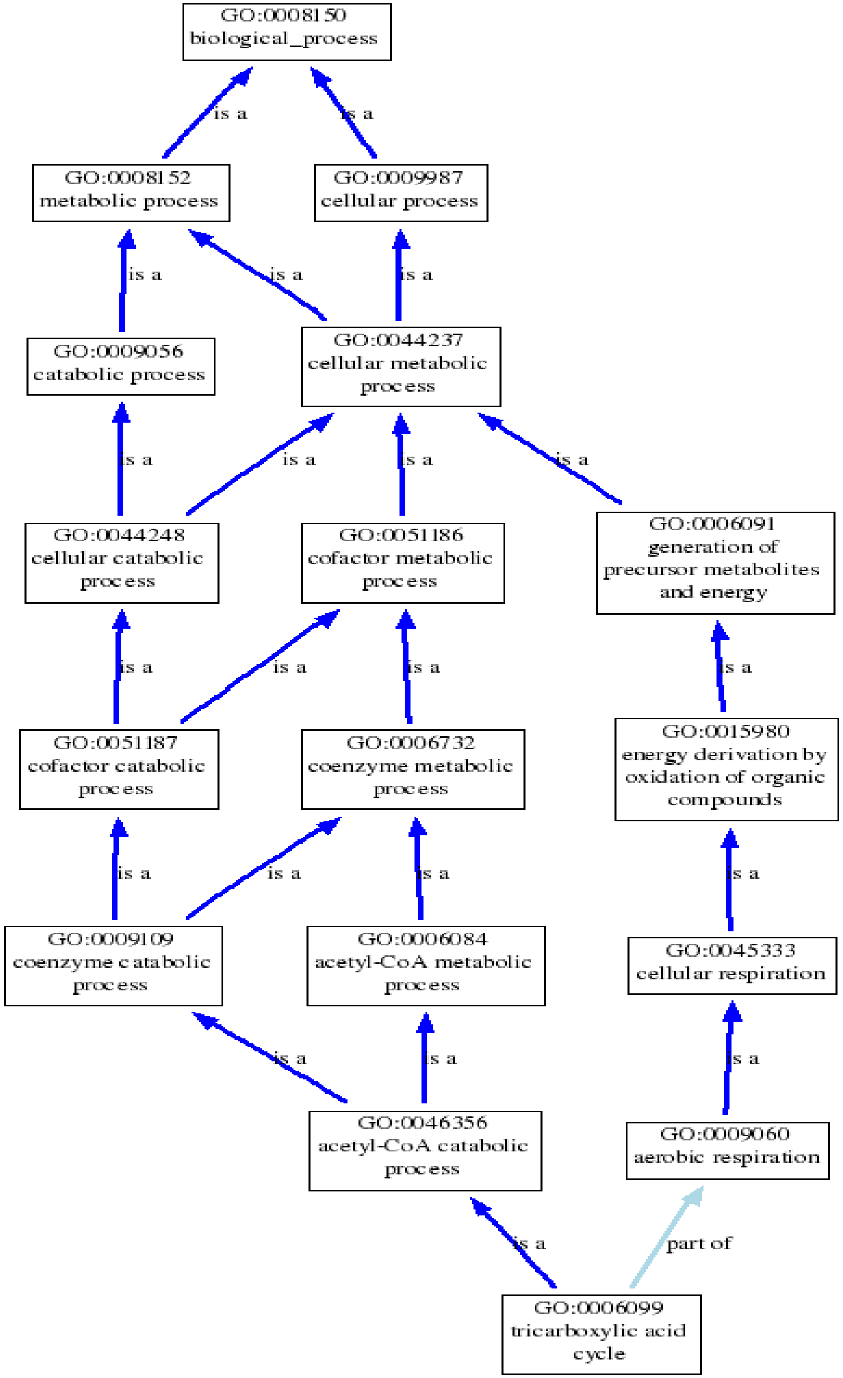
- *Gene Ontology (GO)*
- *Human Phenotype Ontology (HP)*
- *Mammalian Phenotype Ontology (MP)*
- *Merged Disease vocabulary – MEDIC (OMIM → MeSH)*
- *Chemical Entities of Biological Interest (ChEBI)*
- *Anatomical ontologies (MA, ZFA, XAO)*
- *…*

More at OBO Fundry (The Open Biological and Biomedical Ontologies):
http://www.obofoundry.org/

A lot of biological applications, e.g. Functional enrichment (*Subramanian et al* 2005) and semantic similarity(*Yang et al*, 2012).

Gene Ontology

(Ashburner et al., 2000)



GO:0008150
biological_process

GO:0008152
metabolic process

GO:0009987
cellular process

GO:0009056
catabolic process

GO:0044237
cellular metabolic process

GO:0044248
cellular catabolic process

GO:0051186
cofactor metabolic process

GO:0006091
generation of precursor metabolites and energy

GO:0051187
cofactor catabolic process

GO:0006732
coenzyme metabolic process

GO:0015980
energy derivation by oxidation of organic compounds

GO:0009109
coenzyme catabolic process

GO:0006084
acetyl-CoA metabolic process

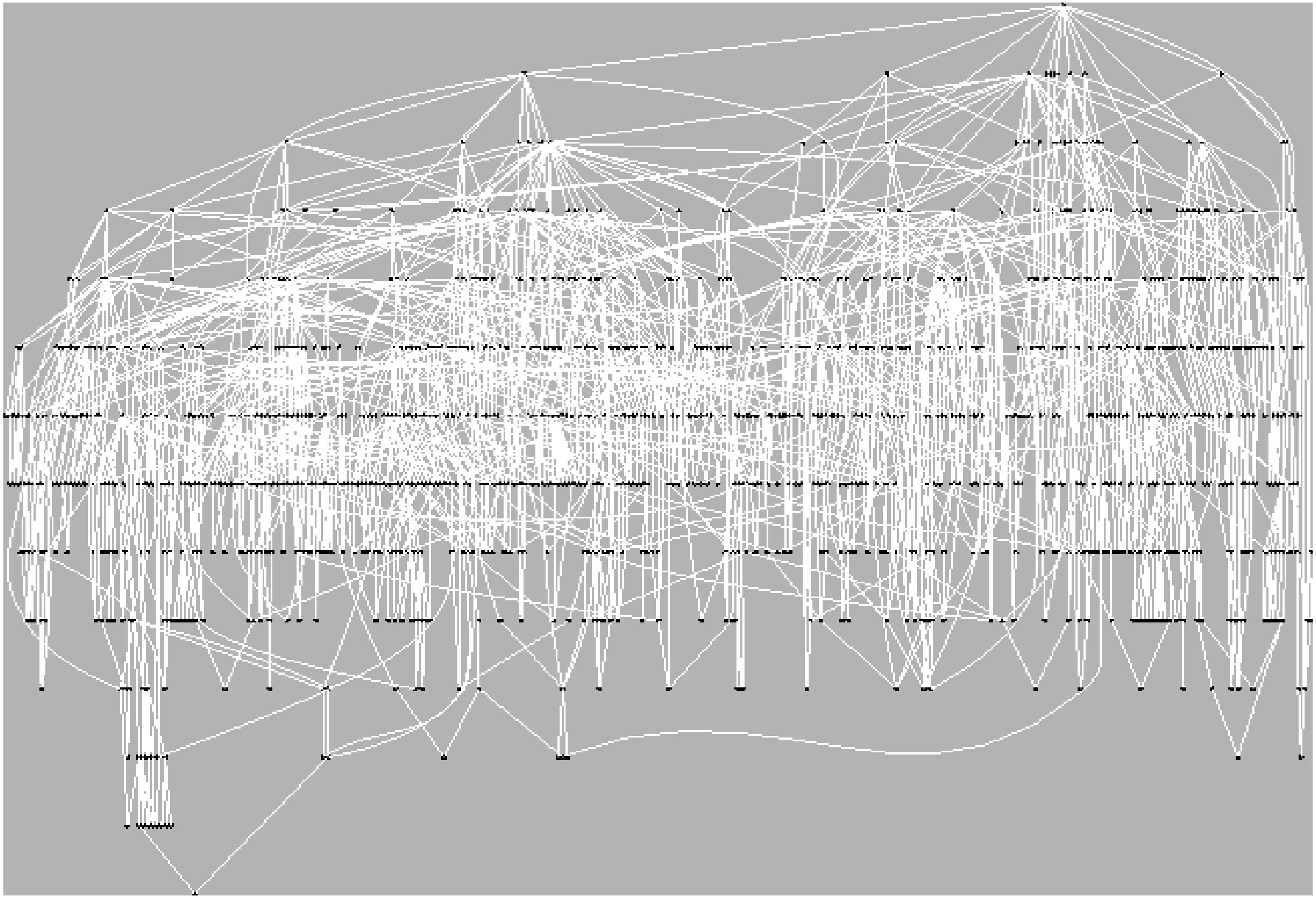
GO:0045333
cellular respiration

GO:0046356
acetyl-CoA catabolic process

GO:0009060
aerobic respiration

GO:0006099
tricarboxylic acid cycle

GO DAG of the BP ontology *(S. cerevisiae)*

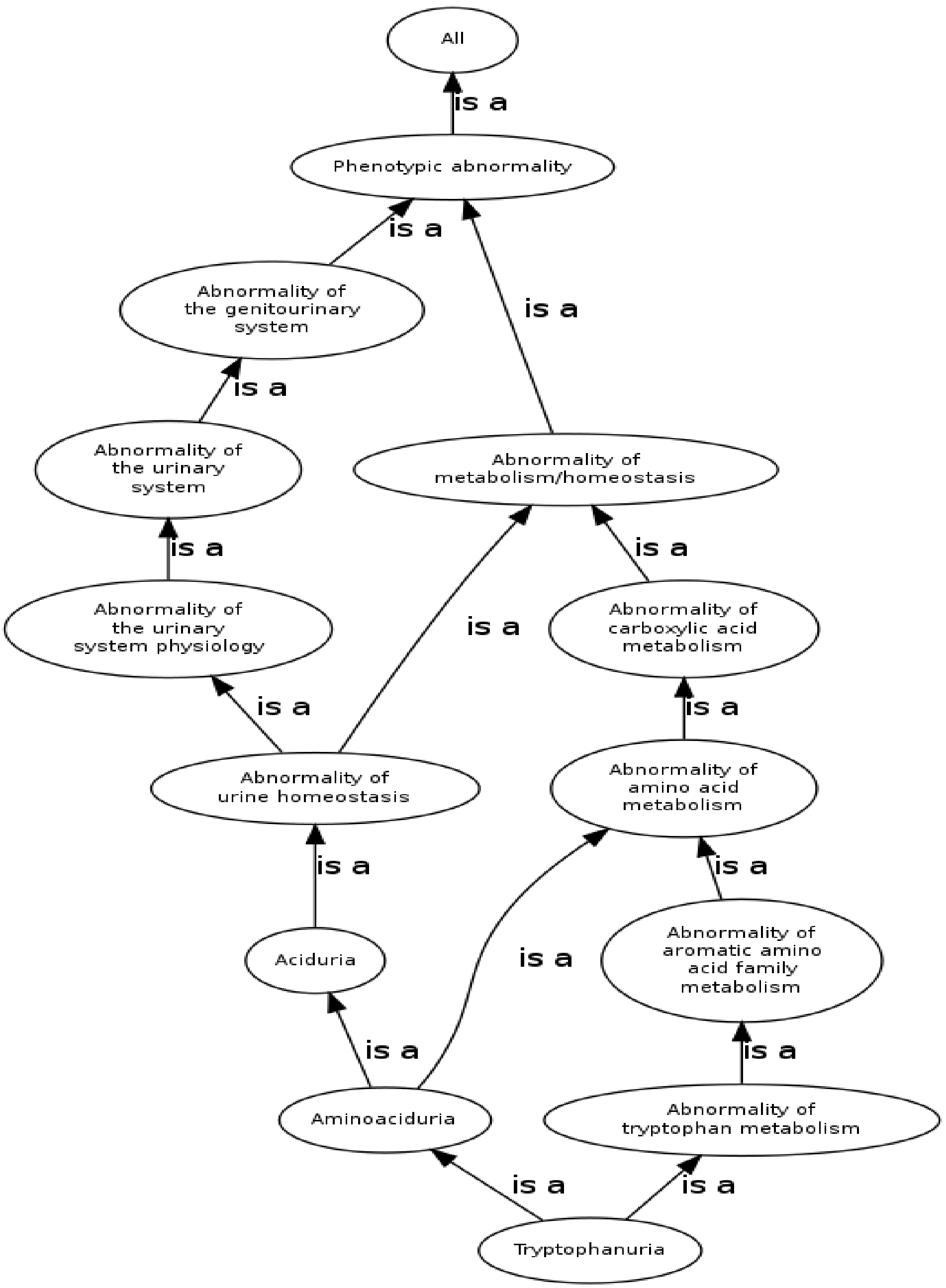


1074 GO classes (nodes) connected by 1804 edges.

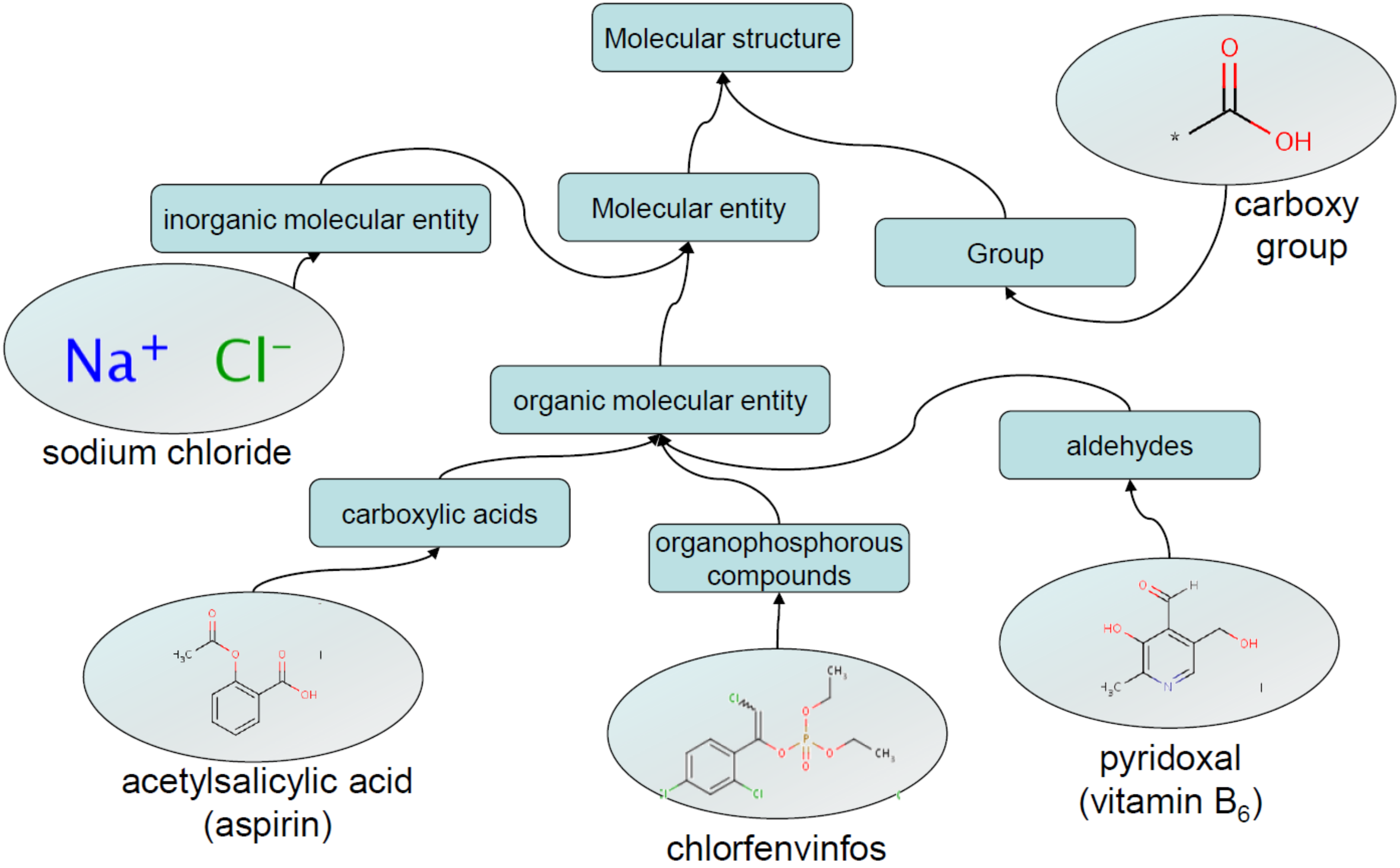
Graph realized through *HCGene* (Valentini and Cesa-Bianchi, 2008)

The Human Phenotype Ontology

(Kohler et al., 2014)


All
is a
Phenotypic abnormality
is a
is a
Abnormality of the genitourinary system
is a
Abnormality of the urinary system
Abnormality of metabolism/homeostasis
is a
is a
is a
Abnormality of the urinary system physiology
Abnormality of carboxylic acid metabolism
is a
is a
Abnormality of urine homeostasis
Abnormality of amino acid metabolism
is a
is a
Aciduria
is a
Abnormality of aromatic amino acid family metabolism
is a
is a
Aminoaciduria
Abnormality of tryptophan metabolism
is a
is a
Tryptophanuria

Molecular structure (sub)ontology (ChEBI)



Classification problems in the context of biological ontologies

- *Ontologies provide predefined taxonomies for several relevant computational problems, e.g.:*
  - Protein Function prediction (GO)
  - Prediction of human gene – abnormal phenotype associations (HP)
  - Prediction of the biological role of small molecules (ChEBI)

⇓

- *Can we design computational methods able to exploit the hierarchical and/or the semantic relationships between ontology terms to provide more robust and accurate predictions?*

AFP is a complex prediction problem characterized by several issues:

- Different level of evidence for functional annotations: *labels at different level of reliability*
- Class frequencies are unbalanced, with positive examples usually largely lower than negatives: *unbalanced classification*
- The notion of “negative example” is not univocally determined: *different strategies to choose negative examples*
- Construction, selection and normalization of the input data are complex and time-consuming: *data preparation is as relevant as the design of the prediction algorithms*
- Multiple sources of data available: *data integration methods*
- Data are usually complex and labels incomplete: *classification with complex and incomplete data*
- Large number of functional classes*: large multi-class classification*
- Multiple annotations for each gene: *multilabel classification*
- Hierarchical relationships between functional classes: *structured multi-label classification*

AFP is a complex prediction problem characterized by several issues:

- Different level of evidence for functional annotations: *labels at different level of reliability*
- Class frequencies are unbalanced, with positive examples usually largely lower than negatives: *unbalanced classification*
- Th
  dif
- Mu
- Co
  co *the*
  des

Can we *design efficient computational methods able to exploit the hierarchical relationships between classes?*

- Data are usually complex and labels incomplete: *classification with complex and incomplete data*
- Large number of functional classes*: large multi-class classification*
- Multiple annotations for each gene: *multilabel classification*
- Hierarchical relationships between functional classes: *structured multi-label classification*

Computational approaches to AFP: a simple taxonomy (*Valentini, 2014*)

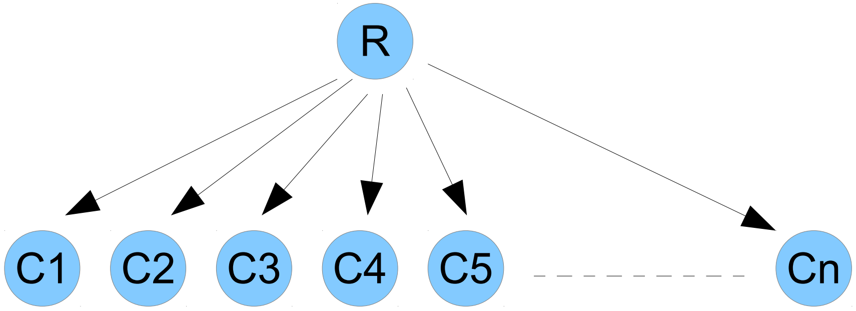
- Inference and annotation transfer through sequence similarity *(Conesa et al 2005, Hamp et al 2013)*
- Network-based methods (*Chua et al, 2007; Mostafavi et al. 2008, Bertoni et al. 2011, Nepusz, Yu and Paccanaro, 2012)*
- Methods based on the joint kernelization of both input and output space (*Astikainen et al. 2008, Sokolov and Ben-Hur, 2010*)
- Hierarchical ensemble methods (*Guan et al. 2006; Obozinski et al, 2008; Schietgat et al. 2010*)

Flat vs hierarchy-aware methods

Flat predictions:

Advantages: simplicity

Drawbacks:
- Inconsitency
- Information loss

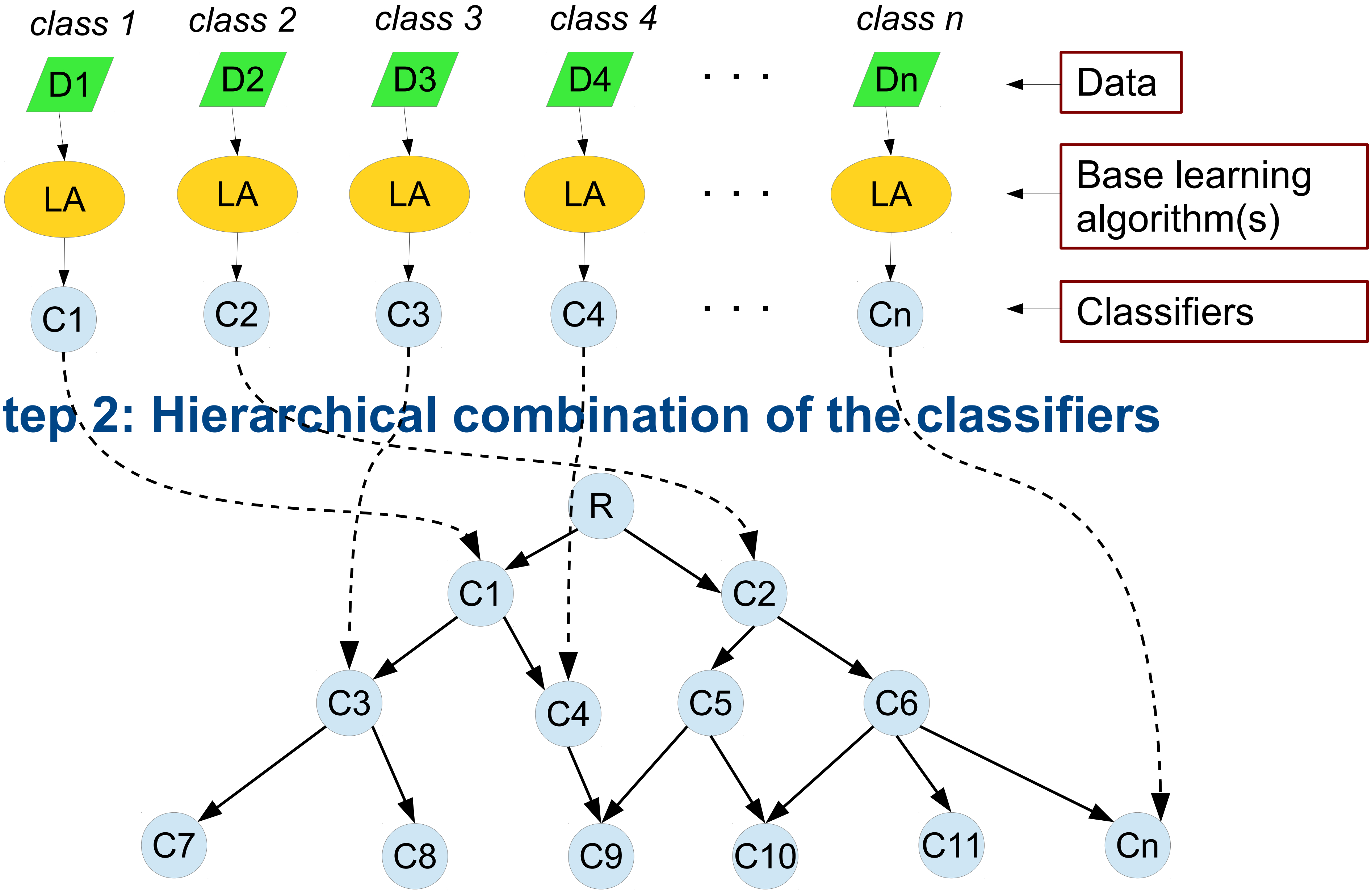


Hierarchy aware approaches:

- Kernel-based structured output methods
  (*Rousu et al. 2006, Sokolov and Ben Hur 2013, Cortes et al. 2014*)
- Hierarchical ensemble methods
  (*Obozinski et al. 2008, Cesa-Bianchi et al. 2012, Yu et al. 2014*)

Hierarchical ensemble methods

Step 1: Training of the base classifiers



Step 2: Hierarchical combination of the classifiers

State-of-the-art Hierarchical ensemble methods

- Most ensembles are conceived for tree-structured taxonomies.
  (Valentini 2011, Cerri et al. 2011, Paes et al 2012, Cesa-Bianchi et al 2012, Hernandez et al 2013)

- Only a few for DAG-structured taxonomies.
  (*Guan et al 2008, Schietgat et al 2010, Yu et al 2015)*

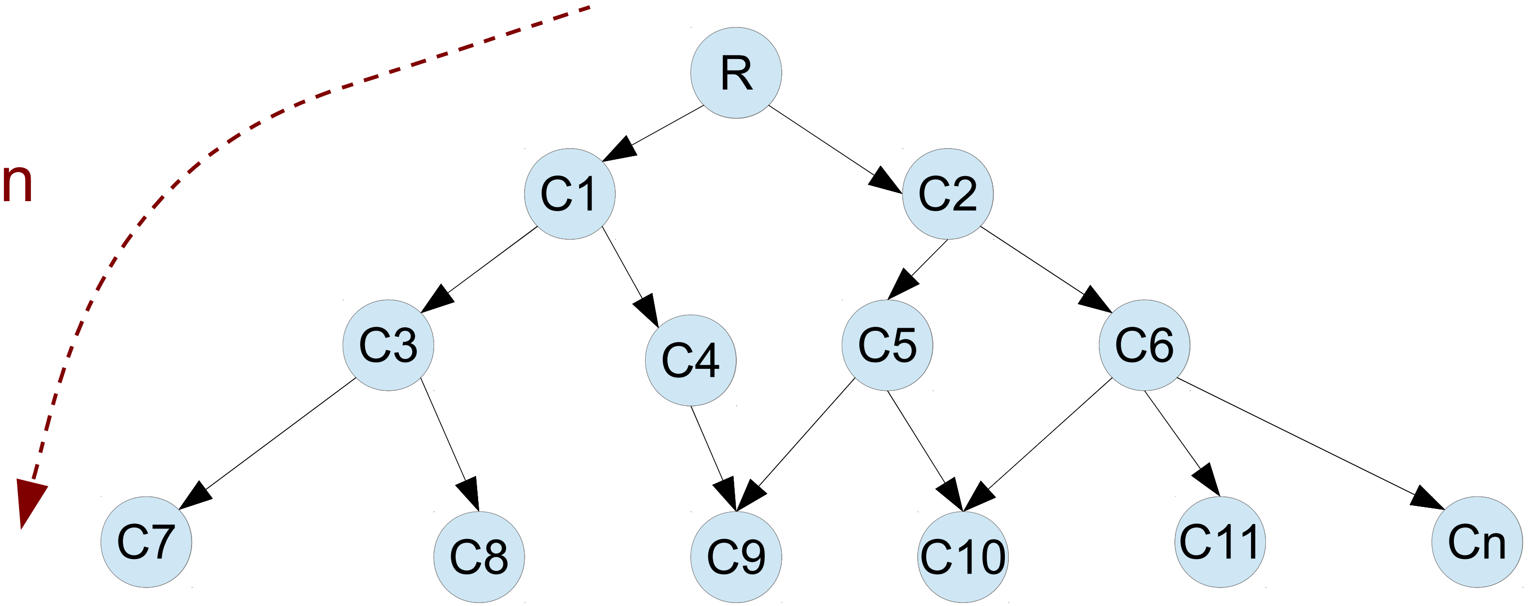
- With DAG-structured taxonomies it is difficult to achieve results comparable with flat methods
  (*Obozinski et al 2008*)

- DAGs are more complex:
  - More parents
  - More edges
  - Multiple paths
  - Nodes may belong to multiple "levels"

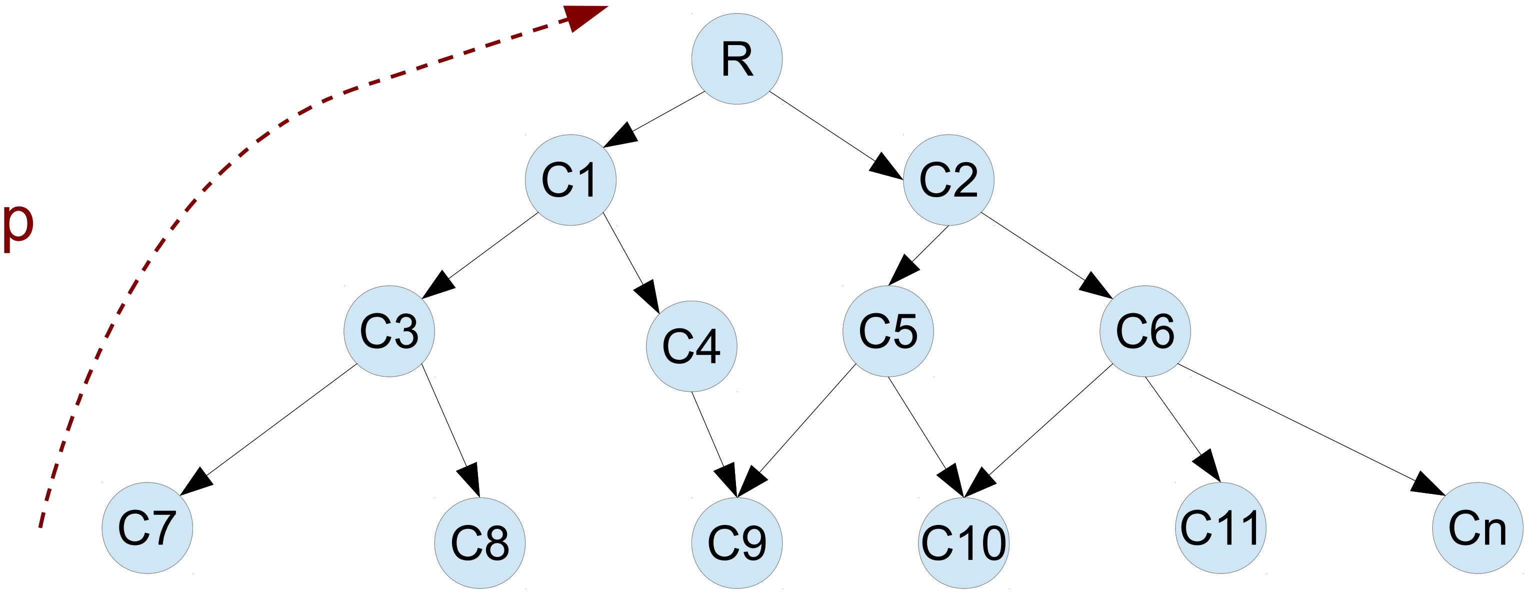
For a recent review on Hierarchical ensemble methods in computational biology, see *Valentini, 2014*

Two general approaches for hierarchical predictions in DAGs

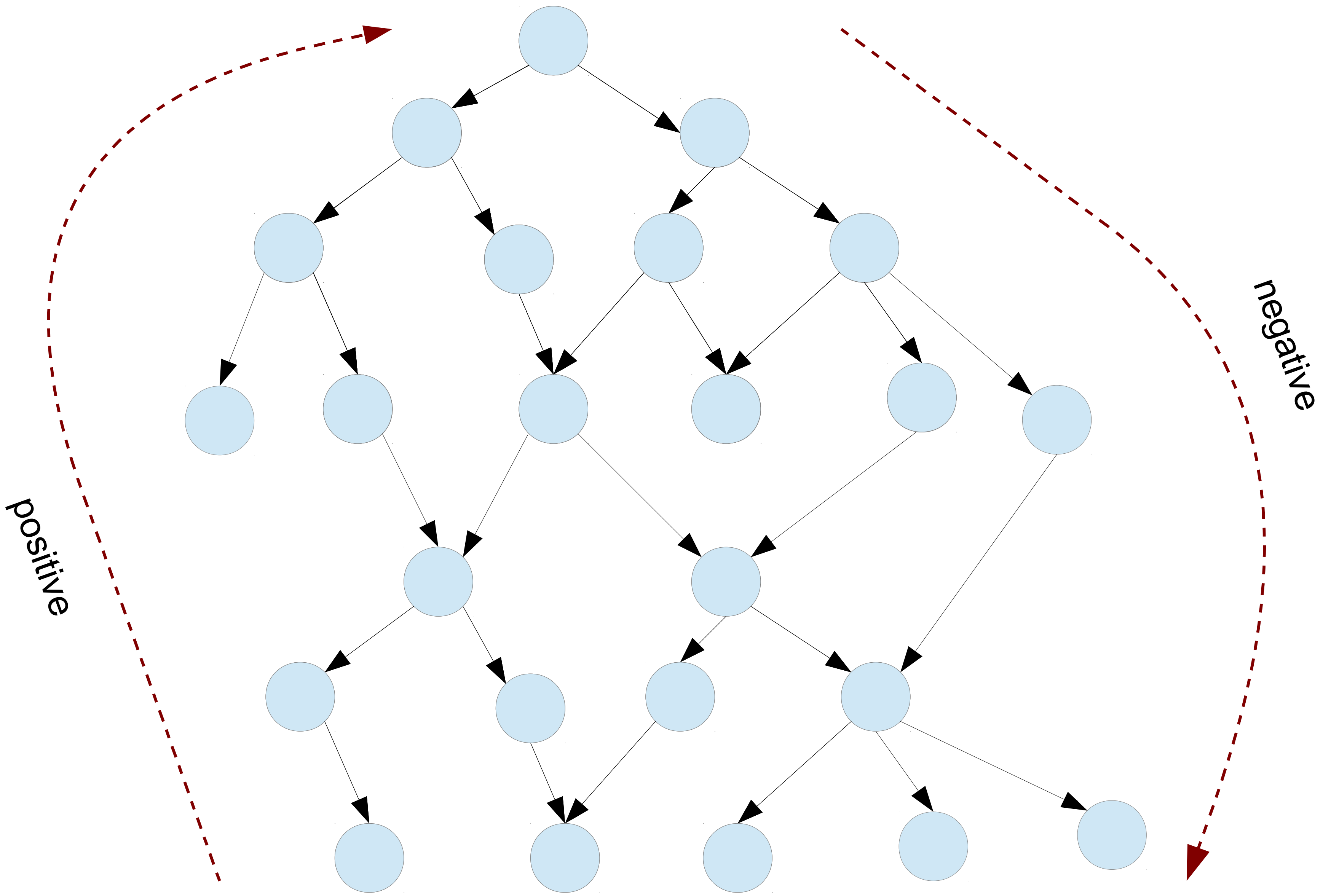
1) Top-Down



2) Bottom-up

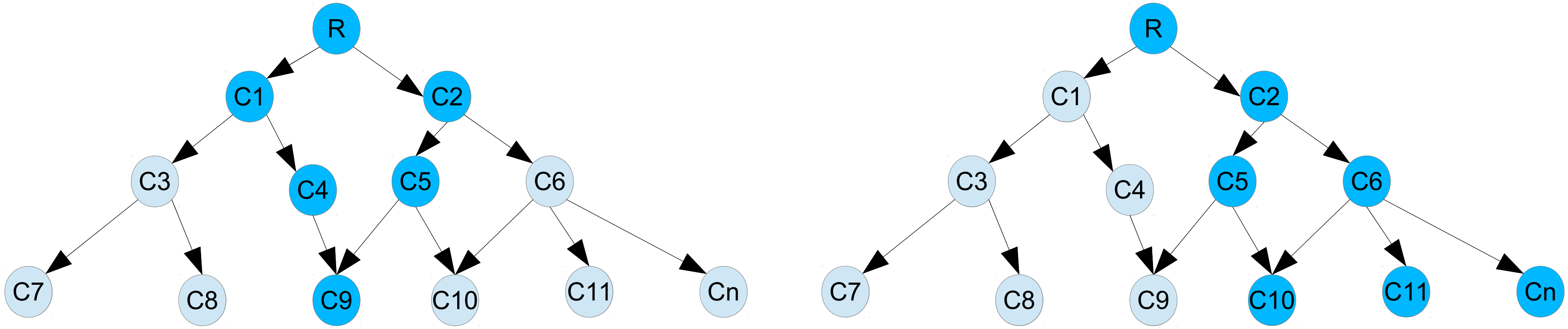


Hierarchical ensemble methods

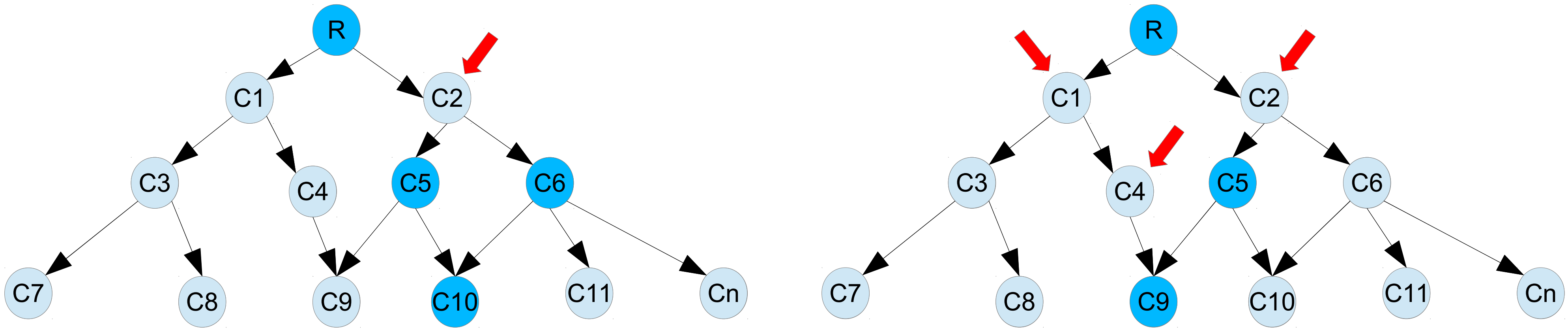


Consistent and inconsistent predictions

Consistent predictions:

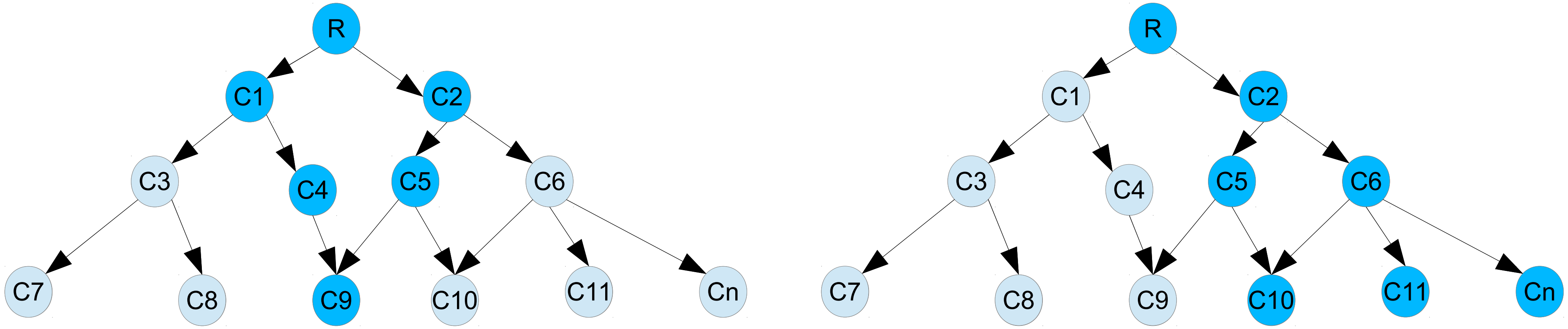


Inconsistent predictions:

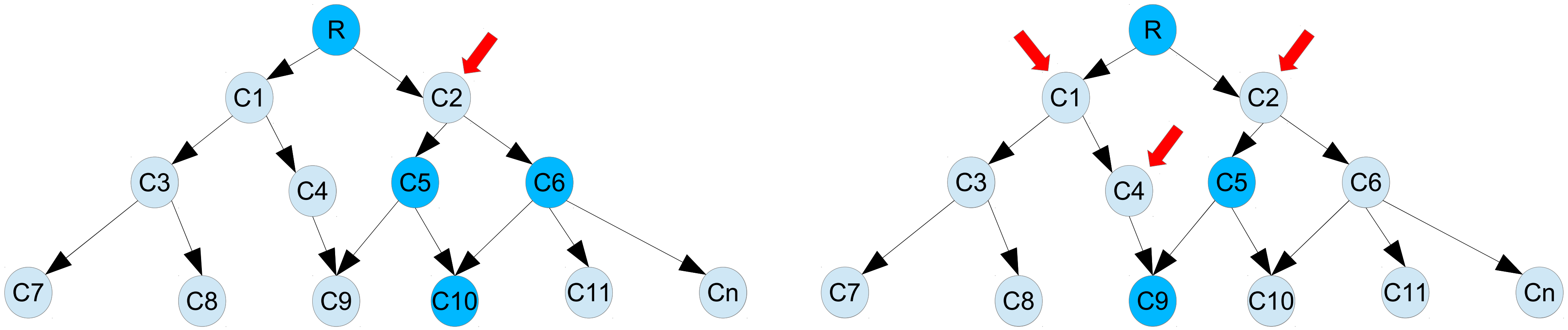


Consistent and inconsistent predictions

Consistent predictions:

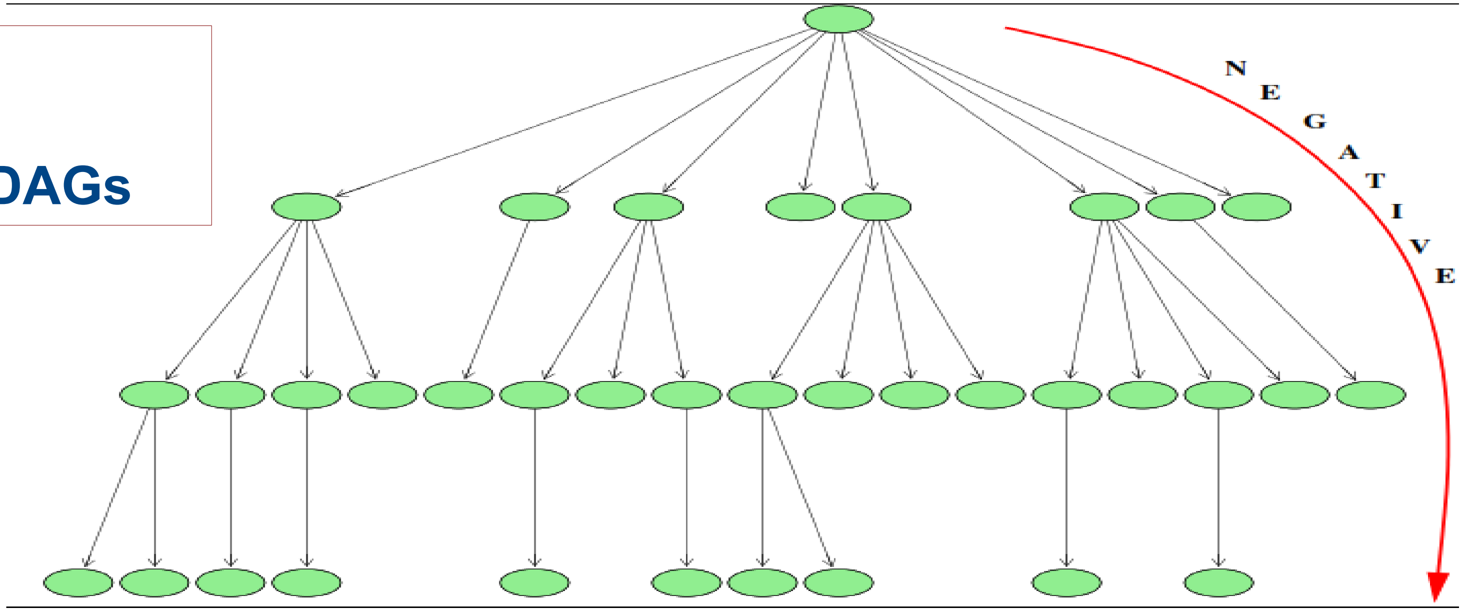


Inconsistent predictions:

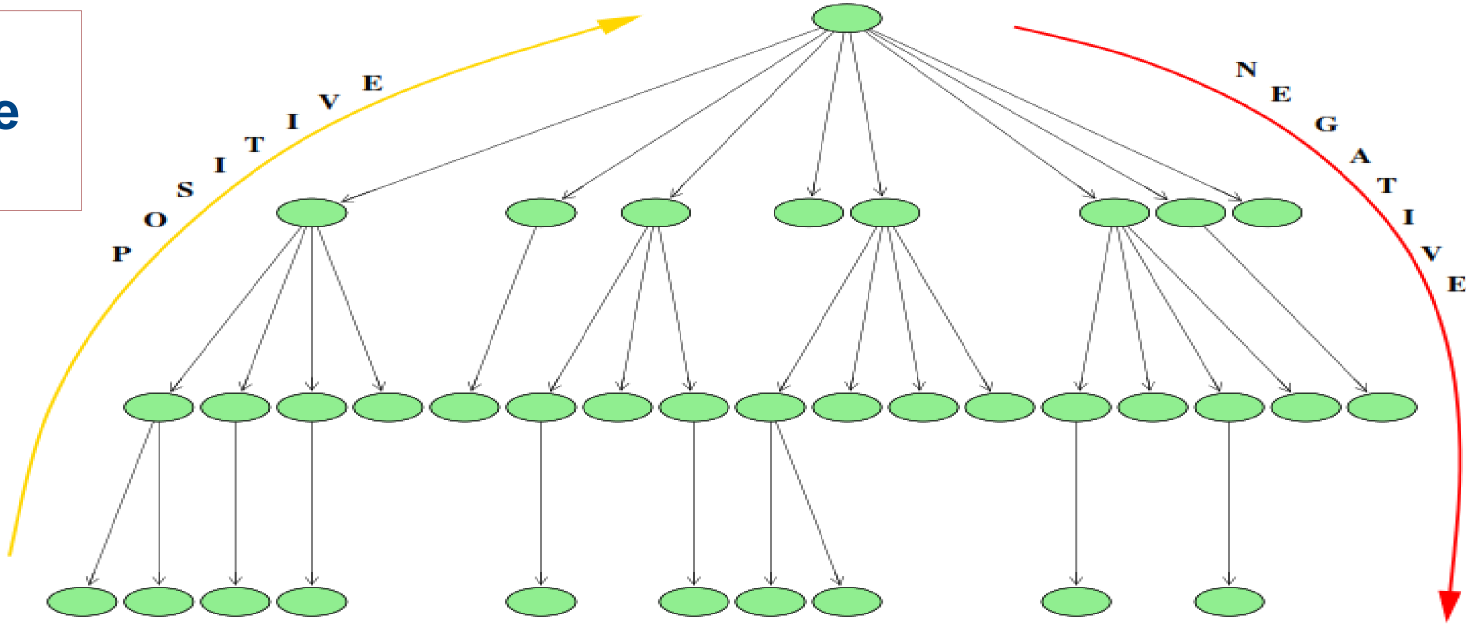


Our proposed approaches

HTD-DAG:
Hierarchical
Top-Down for DAGs



TPR-DAG:
True Path Rule
for DAGs



HTD-DAG – Hierarchical Top-Down ensembles for DAG

A simple rule orderly applied to each class/node:

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

$\hat{y}_i \in [0,1]$ are the flat scores for the class i computed by the base classifier

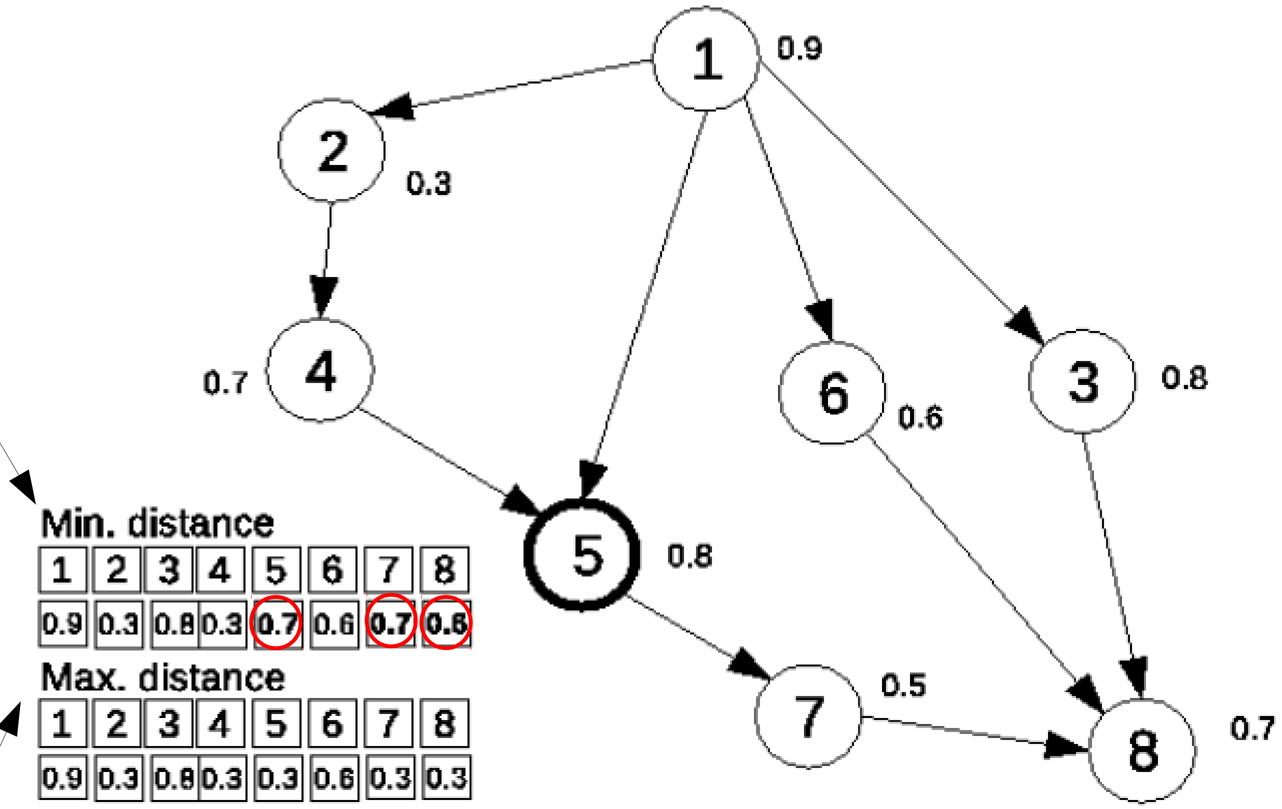
Nodes are processed by level (maximum path length from the root) to assure the consistency of the predictions:

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

Levels must be defined according to the maximum distance from the root

Levels defined in terms of the minimum distance lead to inconsistent predictions

Levels defined in terms of the maximum distance lead to consistent predictions



Min. distance

1	2	3	4	5	6	7	8
0.9	0.3	0.8	0.3	0.7	0.6	0.7	0.6

Max. distance

1	2	3	4	5	6	7	8
0.9	0.3	0.8	0.3	0.3	0.6	0.3	0.3

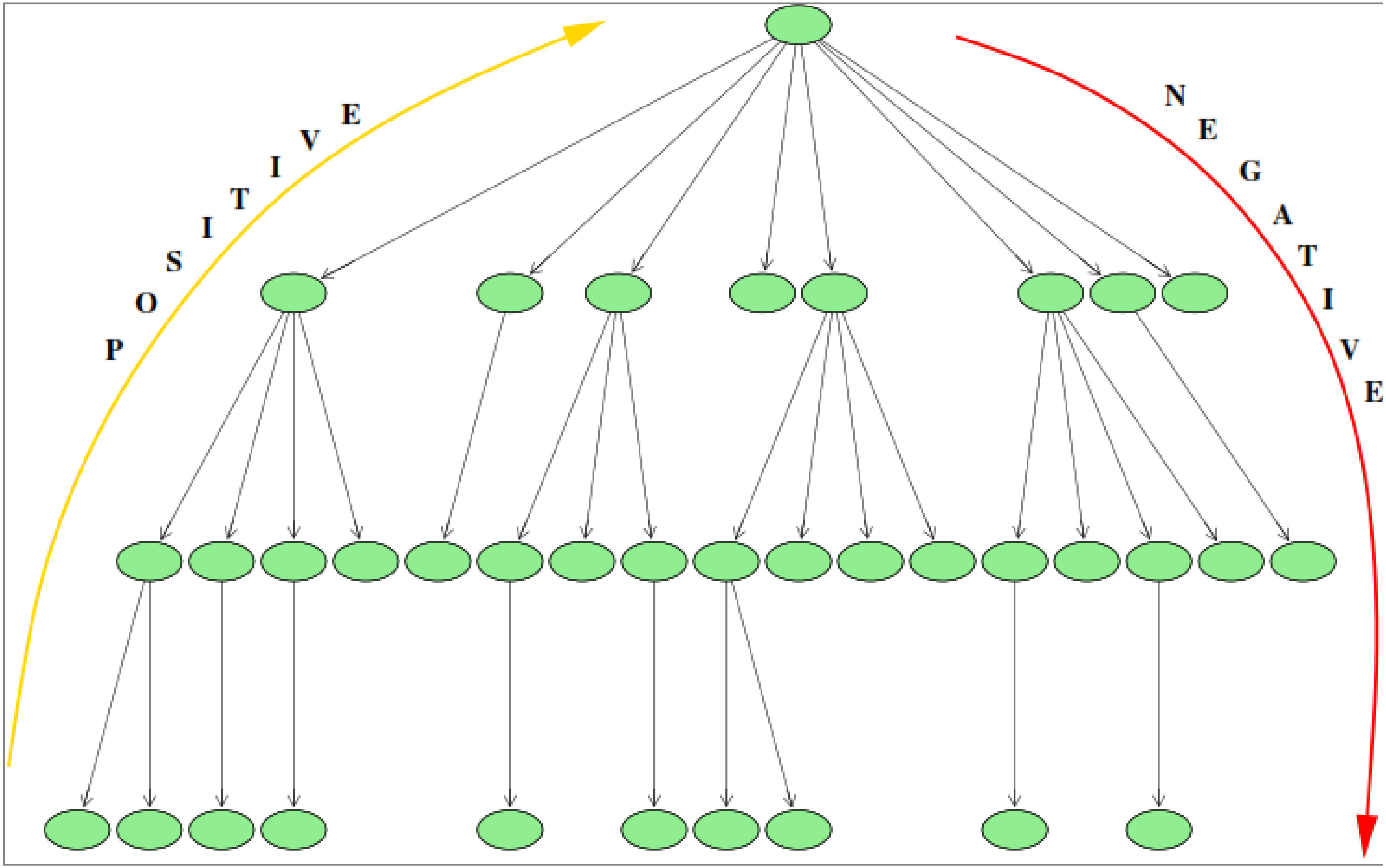
HTD-DAG: the algorithm

HTD-DAG scales linearly with the number of classes

```
Input:
- $G =< V, E >$
- $\hat{\boldsymbol{y}} =< \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|V|} >$ (flat predictions)
begin algorithm
    A. $dist$ := ComputeMaxDist $(G, root(G))$
    B. Per-level top-down visit of $G$:
        $\bar{y}_{root(G)} := \hat{y}_{root(G)}$
        for each $d$ from 1 to $\xi$ do
            $N_d := \{i | dist(i) = d\}$
            for each $i \in N_d$ do
                $x := \min_{j \in par(i)} \bar{y}_j$
                if $(x < \hat{y}_i)$
                    $\bar{y}_i := x$
                else
                    $\bar{y}_i := \hat{y}_i$
            end for
        end for
end algorithm
Output:
- $\bar{\boldsymbol{y}} =< \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{|V|} >$
```

```
ComputeMaxDist (G, r)
begin algorithm
    $s$ = Topological.Sort $(G)$
    $dist[r] := 0$;
    for each $k$ in $V \setminus \{r\}$ do
        $dist[k] := -\infty$
    for each $k$ from 1 to $|V|$ do
        $i := s_k$
        for each $(i, j) \in E$ do
            if $dist[j] < dist[i] + 1$
                $dist[j] := dist[i] + 1$
end algorithm
Output:
- The distance vector $dist$.
```

TPR-DAG: True Path Rule ensembles for DAGs



TPR-DAG: True Path Rule ensembles for DAG

Characterized by a three-step learning strategy:

1. Flat learning of the classes on a per-term basis (a set of independent classification problems)
2. Bottom-up step. Bottom to top propagation of the *positive* predictions → improvement of sensitivity
3. Top-down step. Top to bottom propagation of *negative* predictions → improvement of precision.

Can be considered an adaptation to DAGs of the previously proposed *TPR* algorithm for tree-structured taxonomies (*Valentini*, 2011).

TPR-DAG: Bottom-up step

Flat predictions are modified according to a per-level bottom-up traversal of the DAG:

$$\tilde{y}_i := \frac{1}{1+|\phi_i|}(\hat{y}_i + \sum_{j\in\phi_i} \tilde{y}_j)$$

Where ϕ_i are the "positive" children of i:

$$\phi_i := \{j \in child(i) | \tilde{y}_j > \hat{y}_i\}$$ (Threshold-Free strategy – *TPR-TF*)

$$\phi_i := \{j \in child(i) | \tilde{y}_j > \bar{t}\}$$ (Thresholded strategy – *TPR-T*)

Weighted version of TPR (*TPR-W*):

$$\tilde{y}_i := w\hat{y}_i + \frac{(1-w)}{|\phi_i|}\sum_{j\in\phi_i} \tilde{y}_j$$

TPR-DAG: Top-Down step

A simple rule orderly applied to each class/node (similar to HTD-DAG):

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

$\tilde{y}_i \in [0,1]$ are the scores for the class i computed in the bottom-up step

Nodes are processed by level (maximum path length from the root) to assure the consistency of the predictions:

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

TPR-DAG

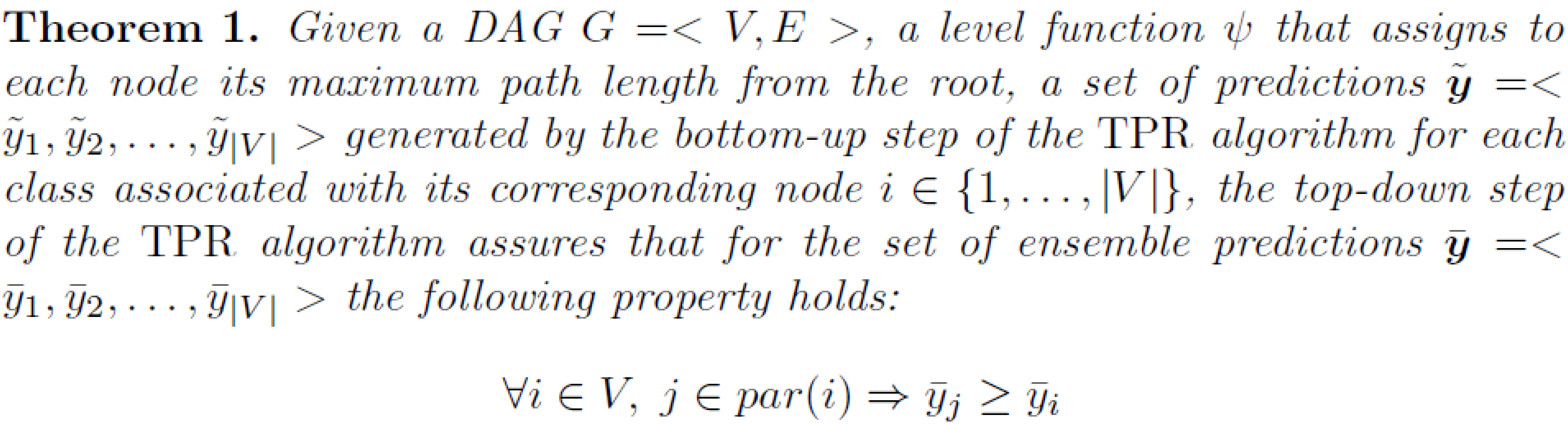
TPR-DAG: the bottom-up and top-down steps

TPR-DAG scales linearly with the number of classes

```
Input:
- $G =< V, E >$
- $V = \{1, 2, \ldots, |V|\}$, 1 is the $root$ node
- $\hat{y} =< \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|V|} >, \quad \hat{y}_i \in [0, 1]$
begin algorithm
    A. Compute $\forall i \in V$ the max distance from $root(G)$:
        $E' := \{e' | e \in E, e' = -e\}$
        $G' :=< V, E' >$
        $dist :=$ Bellman.Ford$(G', root(G'))$
    B. Per-level bottom-up visit of $G$:
        for each $d$ from max$(dist)$ to 0 do
            $N_d := \{i | dist(i) = d\}$
            for each $i \in N_d$ do
                Select $\phi_i$ according to a *positive selection strategy*
                $\tilde{y}_i := \frac{1}{1+|\phi_i|}(\hat{y}_i + \sum_{j \in \phi_i} \tilde{y}_j)$
            end for
        end for
    C. Per-level top-down visit of $G$:
        $\bar{y}_1 := \tilde{y}_1$
        for each $d$ from 1 to max$(dist)$ do
            $N_d := \{i | dist(i) = d\}$
            for each $i \in N_d$ do
                $x := \min_{j \in par(i)} \bar{y}_j$
                if $(x < \tilde{y}_i)$
                    $\bar{y}_i := x$
                else
                    $\bar{y}_i := \tilde{y}_i$
            end for
        end for
end algorithm
Output:
- $\bar{y} =< \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{|V|} >$
```

TPR-DAG and HTD-DAG provide consistent predictions

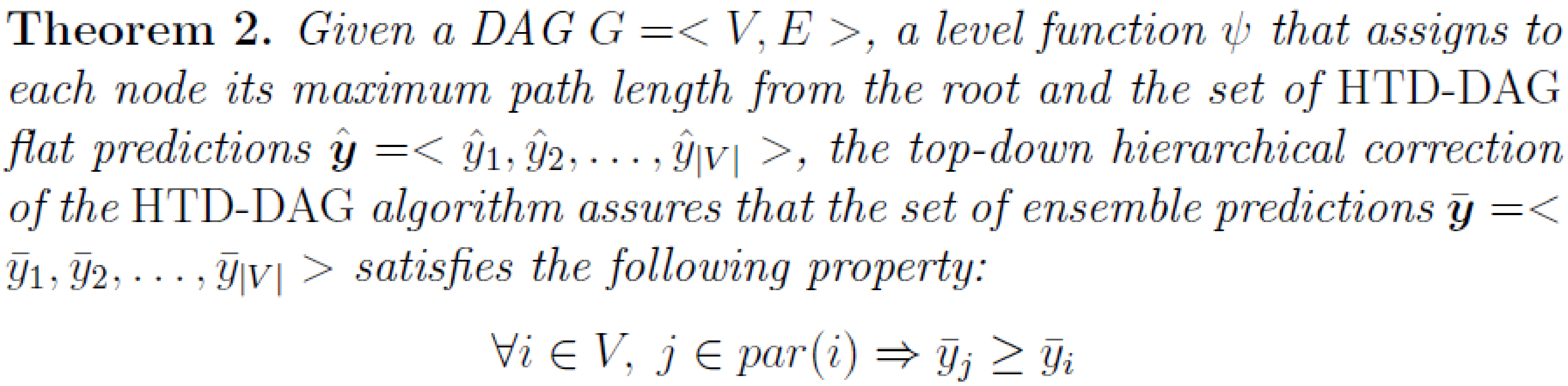
TPR-DAG provides consistent predictions:



Theorem 1. *Given a DAG $G =< V, E >$, a level function ψ that assigns to each node its maximum path length from the root, a set of predictions $\tilde{\boldsymbol{y}} =< \tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{|V|} >$ generated by the bottom-up step of the* TPR *algorithm for each class associated with its corresponding node $i \in \{1, \ldots, |V|\}$, the top-down step of the* TPR *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:*

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

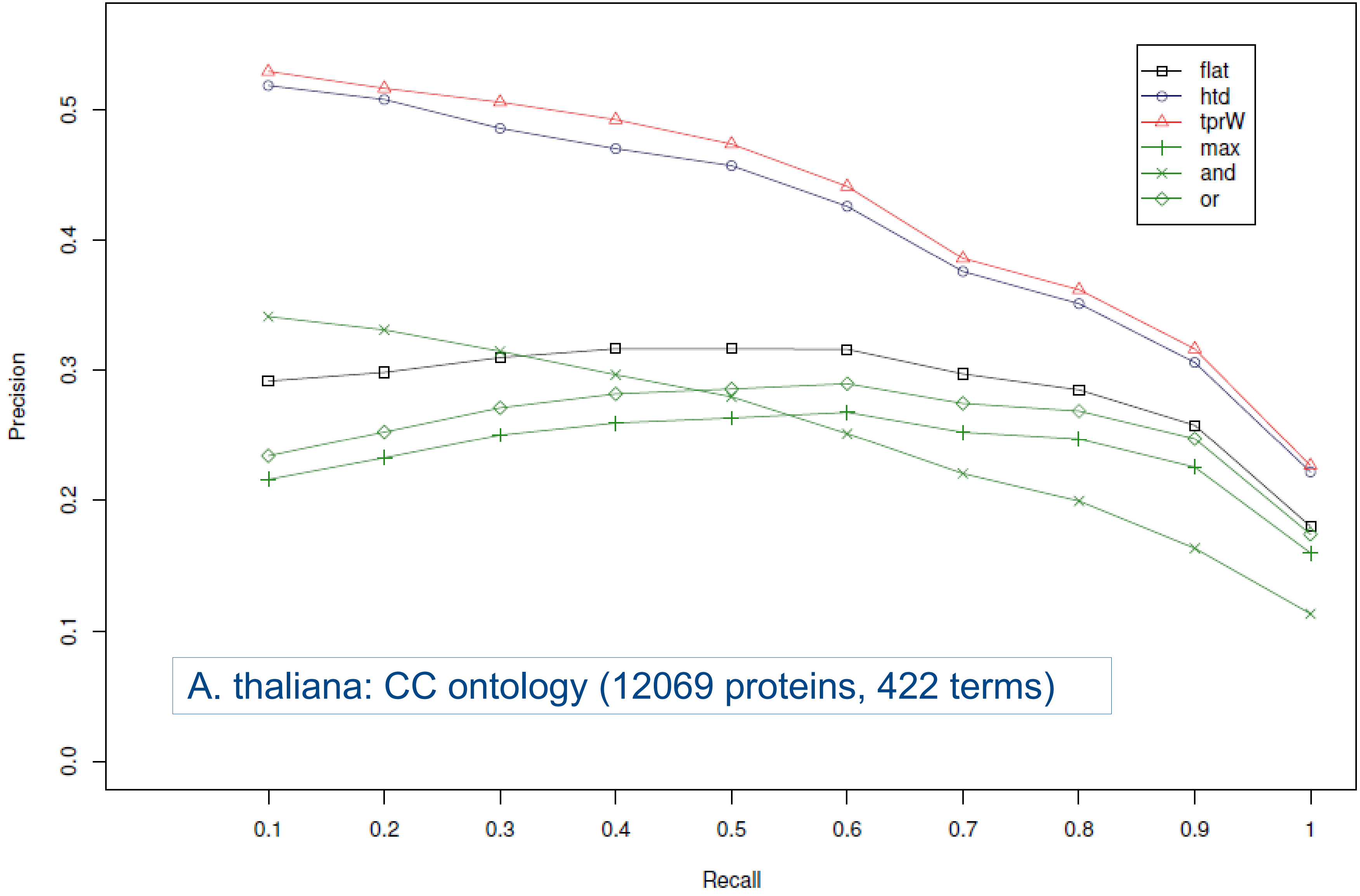
HTD-DAG provides consistent predictions:



Theorem 2. *Given a DAG $G =< V, E >$, a level function ψ that assigns to each node its maximum path length from the root and the set of* HTD-DAG *flat predictions $\hat{\boldsymbol{y}} =< \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|V|} >$, the top-down hierarchical correction of the* HTD-DAG *algorithm assures that the set of ensemble predictions $\bar{\boldsymbol{y}} =< \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{|V|} >$ satisfies the following property:*

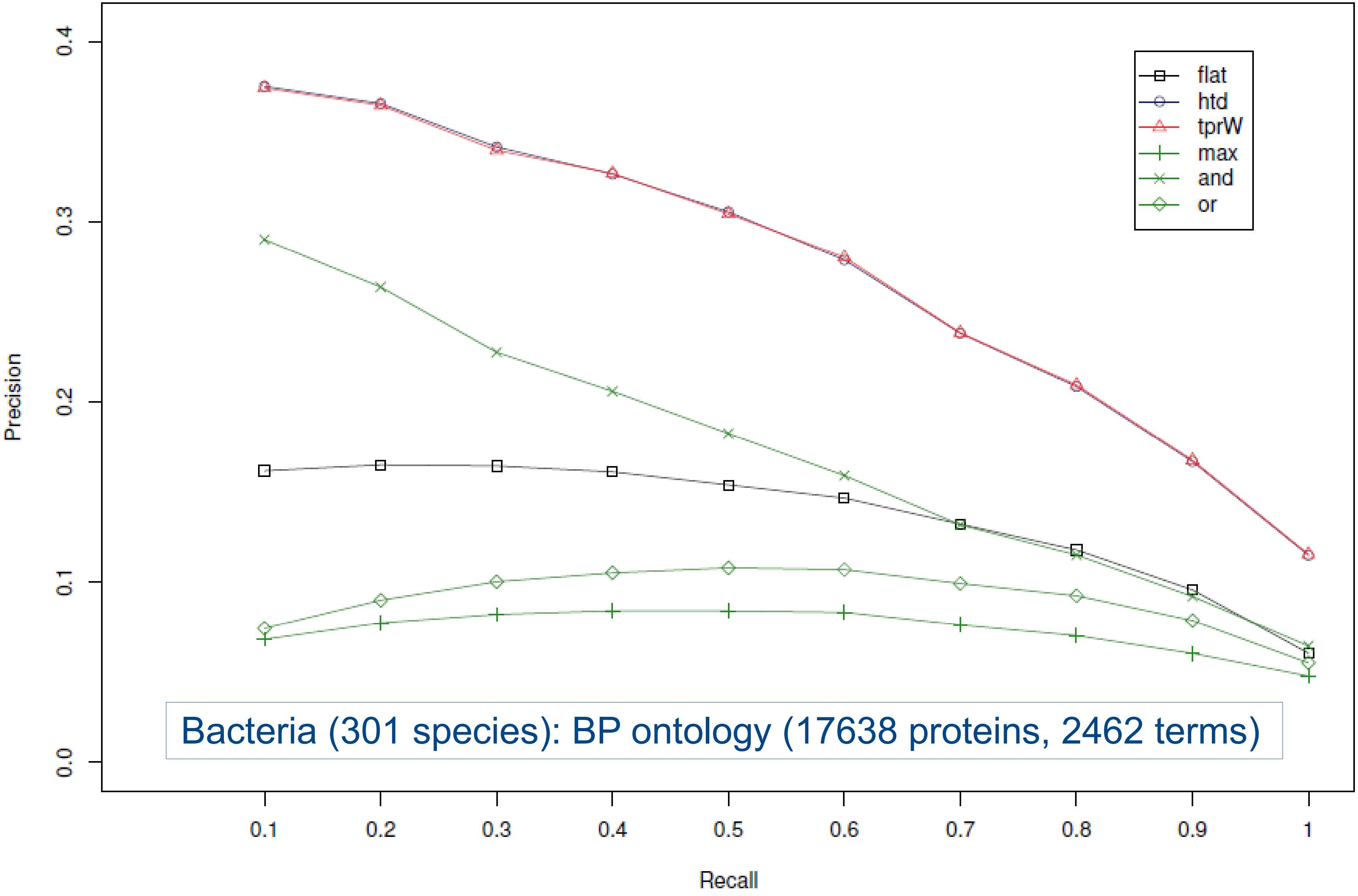
$$\forall i \in V,\ j \in par(i) \Rightarrow \bar{y}_j \geq \bar{y}_i$$

TPR-DAG and HTD-DAG significantly improve flat methods in the protein function prediction problem



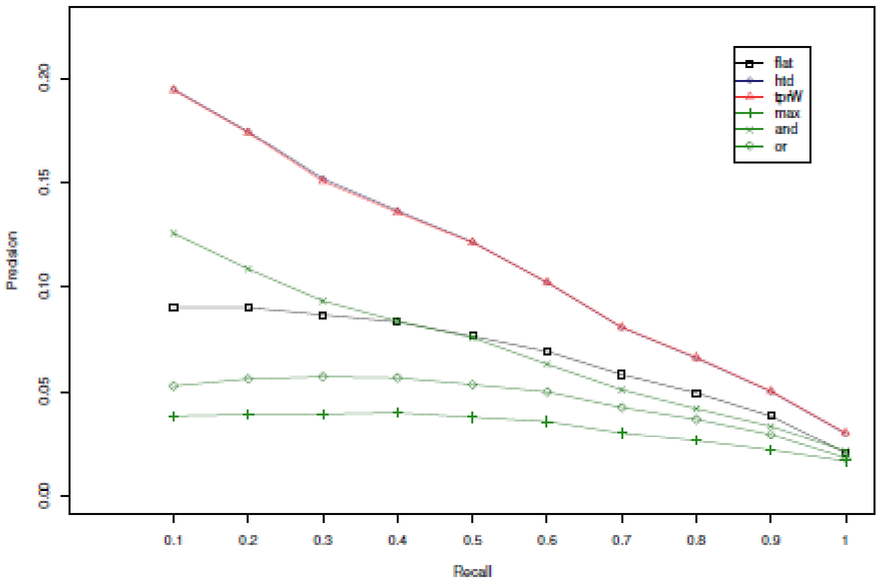
A. thaliana: CC ontology (12069 proteins, 422 terms)

TPR-DAG and HTD-DAG significantly improve flat methods in the protein function prediction problem

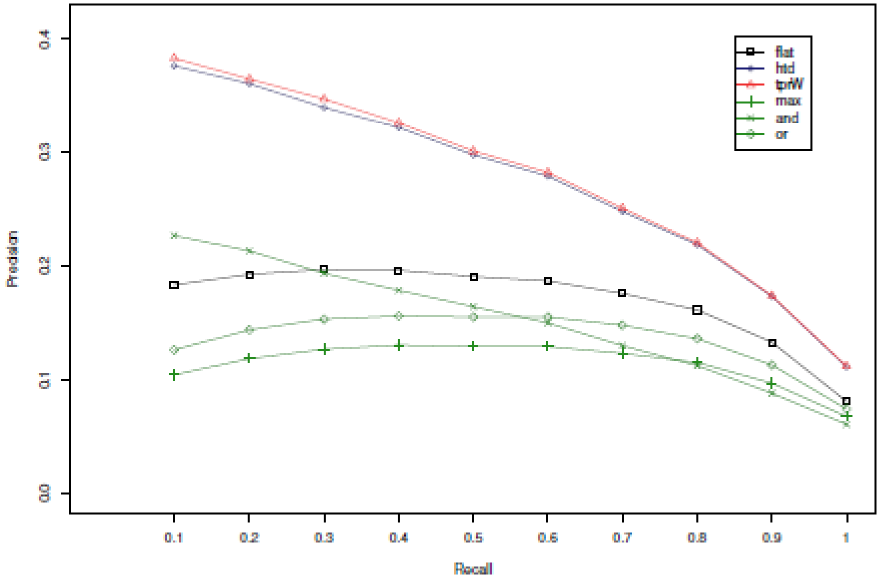


Bacteria (301 species): BP ontology (17638 proteins, 2462 terms)

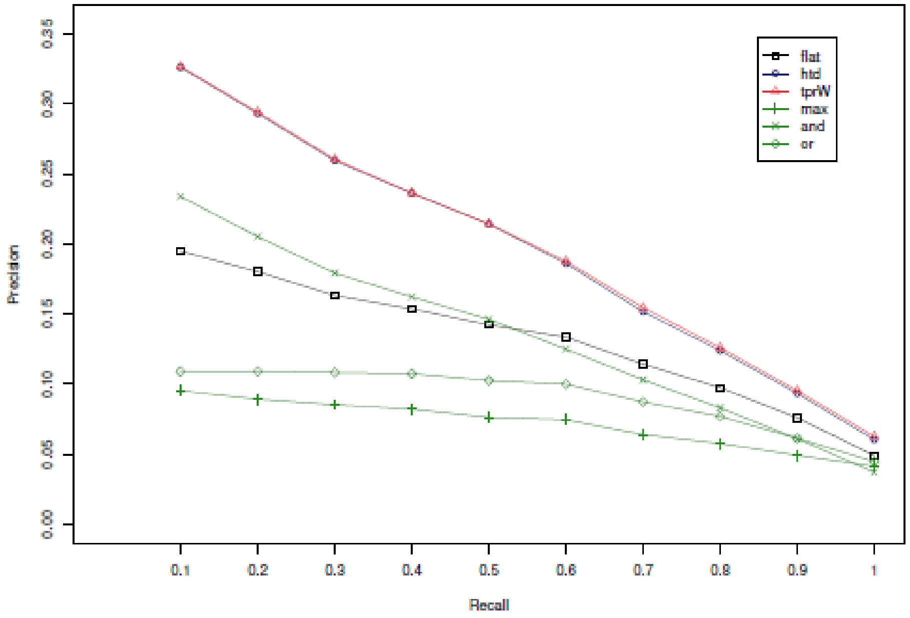
HTD-DAG and TPR-DAG



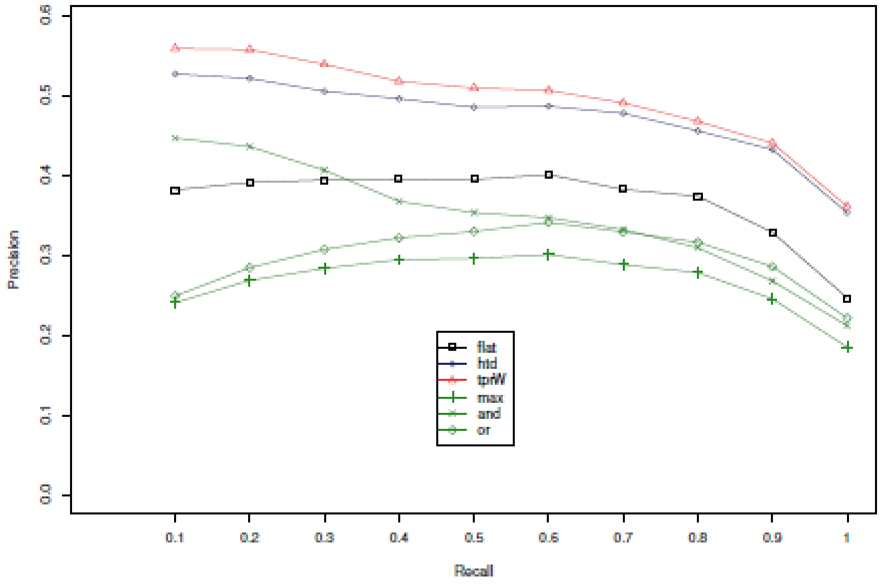
H. *sapiens*: BP ontology (20257 proteins, 8310 terms)



H. *sapiens*: CC ontology (20257 proteins, 961 terms)

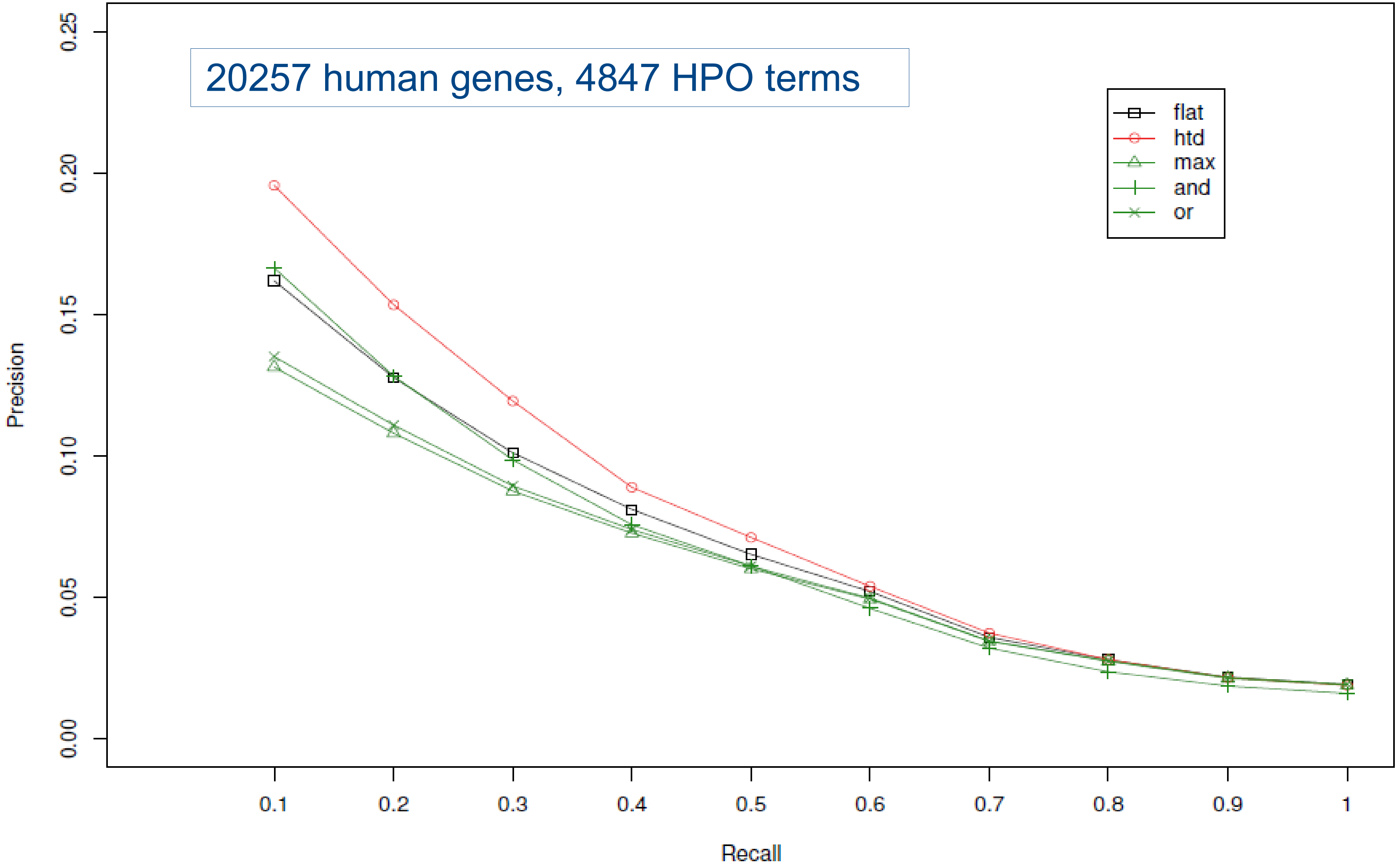


A. *thaliana*: BP ontology (12069 proteins, 3410 terms)



Bacteria (301 species): CC ontology (17638 proteins, 210 terms)

HTD-DAG significantly improves flat methods in the prediction of human gene abnormal phenotype associations



G. Valentini, S. Kohler, M. Re, M. Notaro, P.N. Robinson, **Prediction of human gene – phenotype association by exploiting the hierarchical structure of the Human Phenotype Ontology**, *3rd International Work-Conference on Bioinformatics and Biomedical Engineering* - IWBBIO 2015, *Lecture Notes in Bioinformatics*, vol. 9043, pp. 66-77, Springer (2015)

Scalability of HTD and TPR-DAG

- HTD-DAG and TPR-DAG are both *linear in time* with respect the number of terms (classes) of the ontology
- Each example (protein) can be processed one at a time (or in constant chunks): *(sub)linear complexity in space*

➡

Big ontologies and large number of proteins can be processed with ordinary computers

Example:

On going application to big multi-species protein function prediction problems:

- more than 400 organisms
- 1.5 millions of proteins (core of the STRING database)
- Construction of a multi-species network including hundreds of millions of edges (intra and inter-species)
- Scalable vertex-centric and secondary memory-based computation
- Thousands of GO functional classes to be predicted
- Scalable hierarchical correction of the predictions

Conclusions and future developments

- Relevant problems in computational biology can be modeled through hierarchical ontologies

- HTD-DAG and TPR-DAG:
  a) scale linearly
  b) provide consistent predictions
  c) improve flat predictions
  d) can be applied to big data

- Developments and future work:
  a) Are hierarchical ensembles meta-learning tools that can improve any flat approach? → More theoretical insights and experiments with different base learners
  b) Application in the context of complex MAFP problems
  c) TPR-DAG is a family of algorithms: experimenting with new variants
  d) Design of novel TPR algorithms working on the trade-off sensitivity/precision.
  e) Exploiting the hierarchy just in the first step (e.g.: multi-task learning)

Acknowledgments

AnacletoLab:

Matteo Re

Marco Mesiti

Marco Frasca

Jianyi Lin

Marco Notaro

Peter Robinson and Sebastian Kohler (Charité, Humbolt Universitat und Freie Universitat, Berlin)

And thanks also to Anacleto !

http://anacletolab.di.unimi.it

References:

G. Valentini, M. Frasca, M. Re. A Hierarchical Top-Down ensemble method for multi-label prediction in DAG-structured taxonomies (submitted)

G. Valentini, S. Kohler, M. Re, M. Notaro, P.N. Robinson. Prediction of human gene -phenotype associations by exploiting the hierarchical structure of the Human Phenotype Ontology, IWBBIO 2015, *Lecture Notes in Bioinformatics,* vol.9043, pp.66-77, 2015

M. Mesiti, M. Re, G. Valentini.Think globally and solve locally: secondary memory-based network learning for automated multi-species function prediction, *GigaScience*, 3:5, 2014

G. Valentini, Hierarchical Ensemble Methods for Protein Function Prediction, *ISRN Bioinformatics*, vol. 2014 (2014)

N. Cesa-Bianchi, M. Re, G. Valentini. Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference, *Machine Learning*, vol.88(1), pp. 209-241, 2012

G. Valentini. True Path Rule hierarchical ensembles for genome-wide gene function prediction, *IEEE ACM Transactions on Computational Biology and Bioinformatics*, 8(3) pp. 832-847, 2011