Leveraging Sequence Classification by Taxonomy-based Multitask Learning

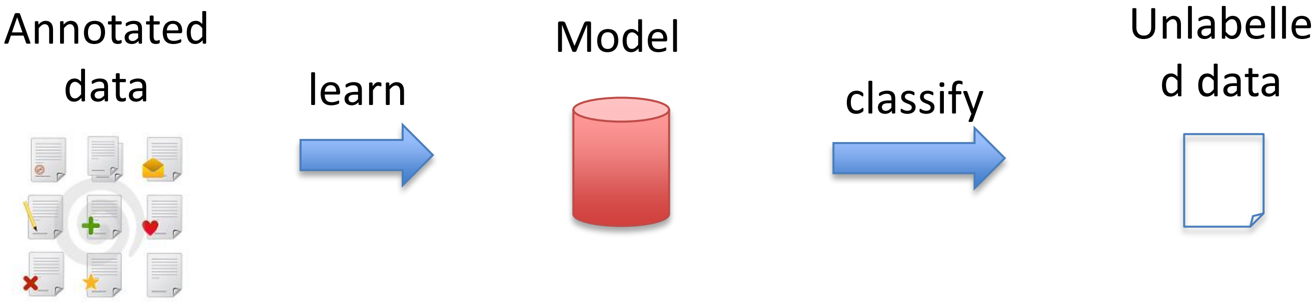
Christian Widmer, Jose Leiva, Yasemin Altun, Gunnar Ratsch

Presented by
Meghana Kshirsagar

Outline

- Multitask Learning setting
- Application:
  - prediction of splicing sites across organisms
- 3 approaches to multi-task learning
  - Top-down
  - Pairwise
  - Multitask kernel
- Experiments and Results

Prelude: Classification



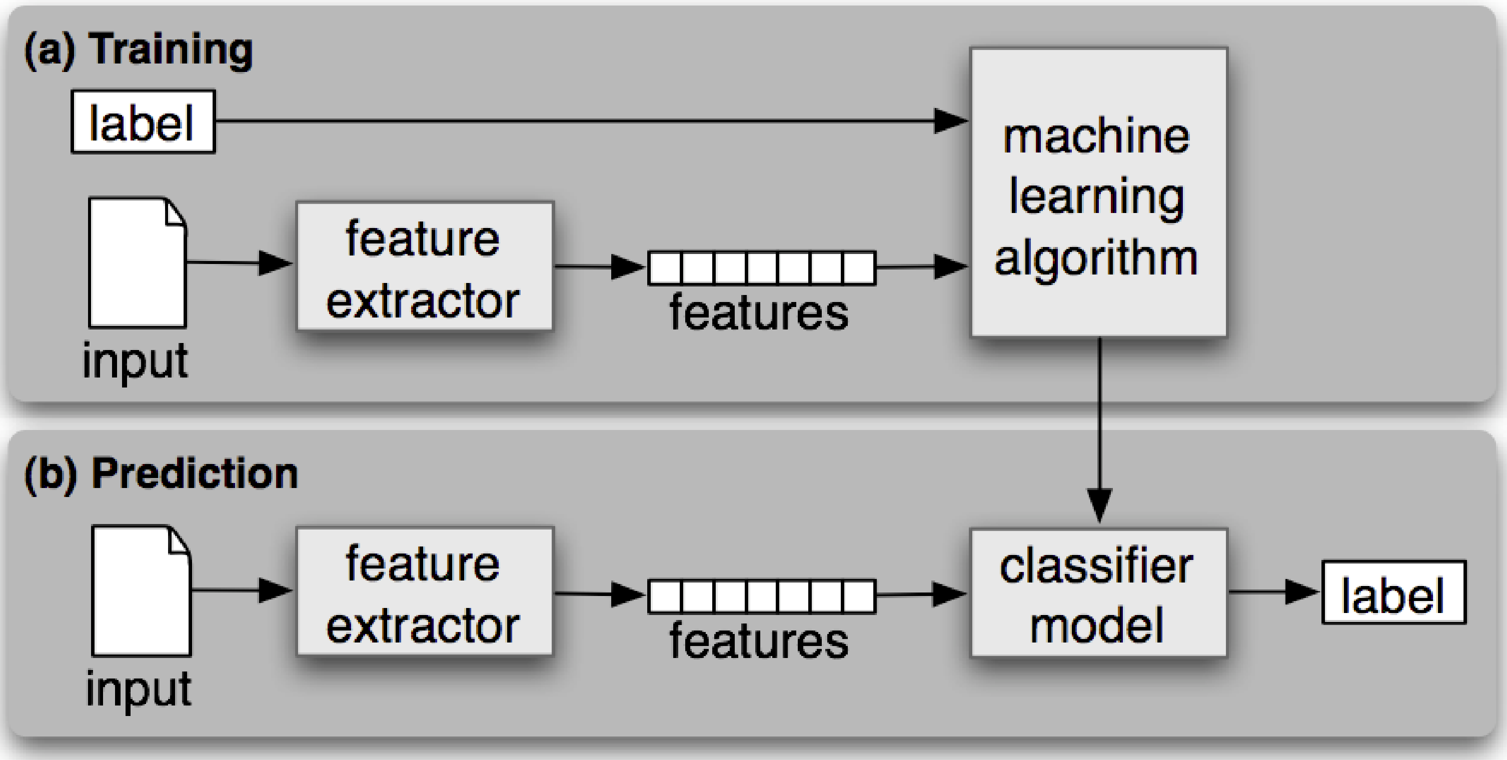
Annotated data

learn

Model

classify

Unlabelled data



Multitask learning setting

Multi-task learning:

- Learning multiple concepts together
  - simultaneously
  - transfer learning between concepts

Why Multi-task learning?

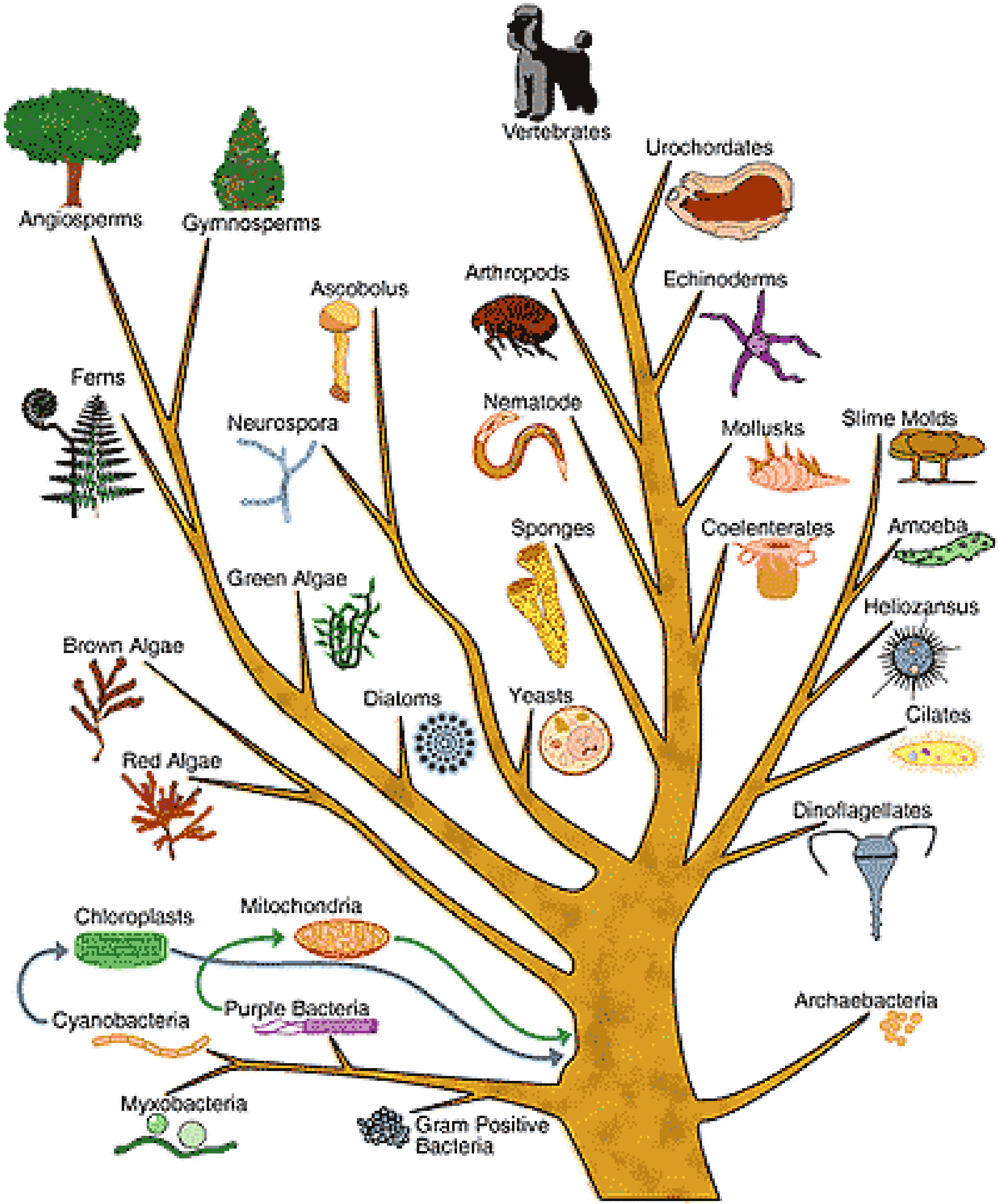
- Model quality limited by insufficient training data
  - exploit similarity between tasks
- In Computational Biology,
  - Organisms share evolutionary history
  - Many biological processes are conserved

Setting

- Consider *M* tasks: T_1, ..., T_M
- We are given data D_1, ..., D_M for each task
  - D_i = { (x_1, y_1), ... , } for each task
- Build *M* classifiers, using all available information from all tasks
- Use a taxonomy to learn these multiple tasks!

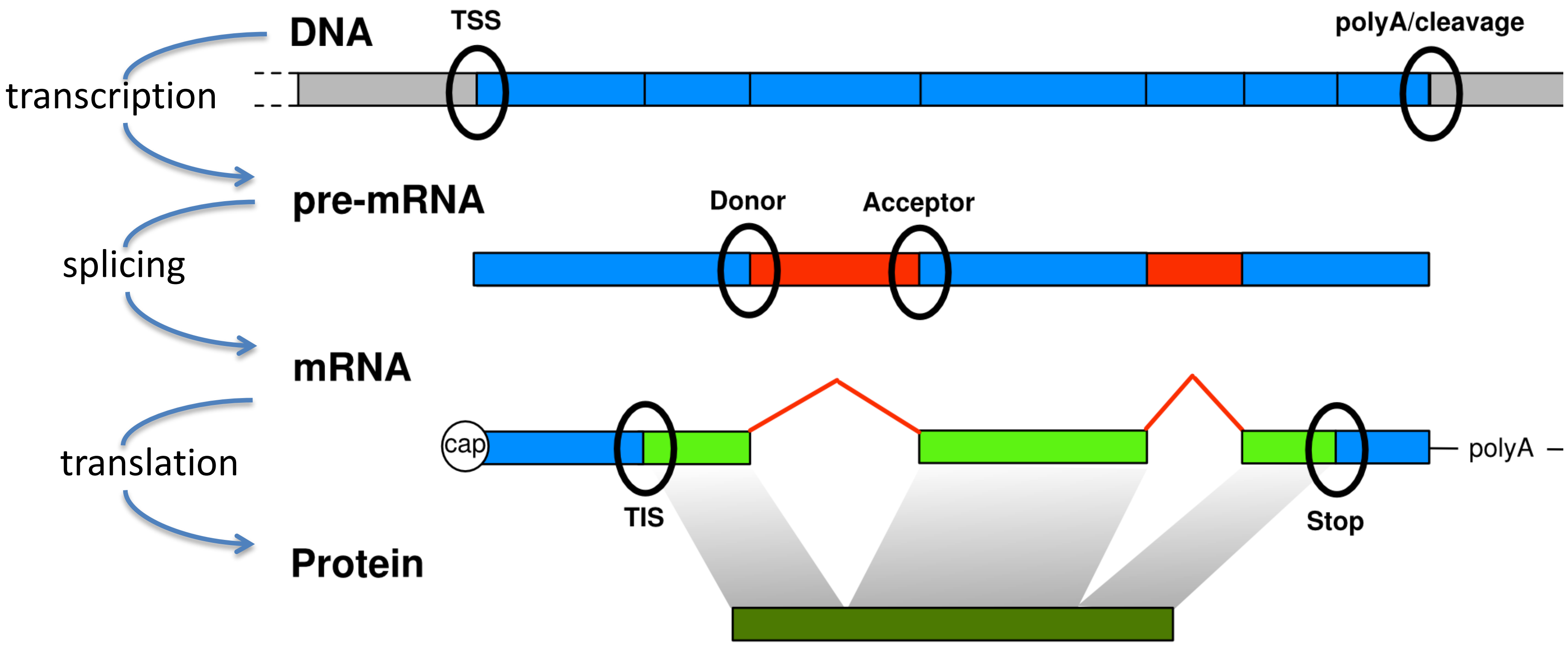
Why use Taxonomy?

- Taxonomy can be used to define relationships between two tasks
- In biology, taxonomy naturally arises from a phylogenic tree
- Closer tasks will benefit more from each other



Application to a biological problem

Prediction of splicing sites

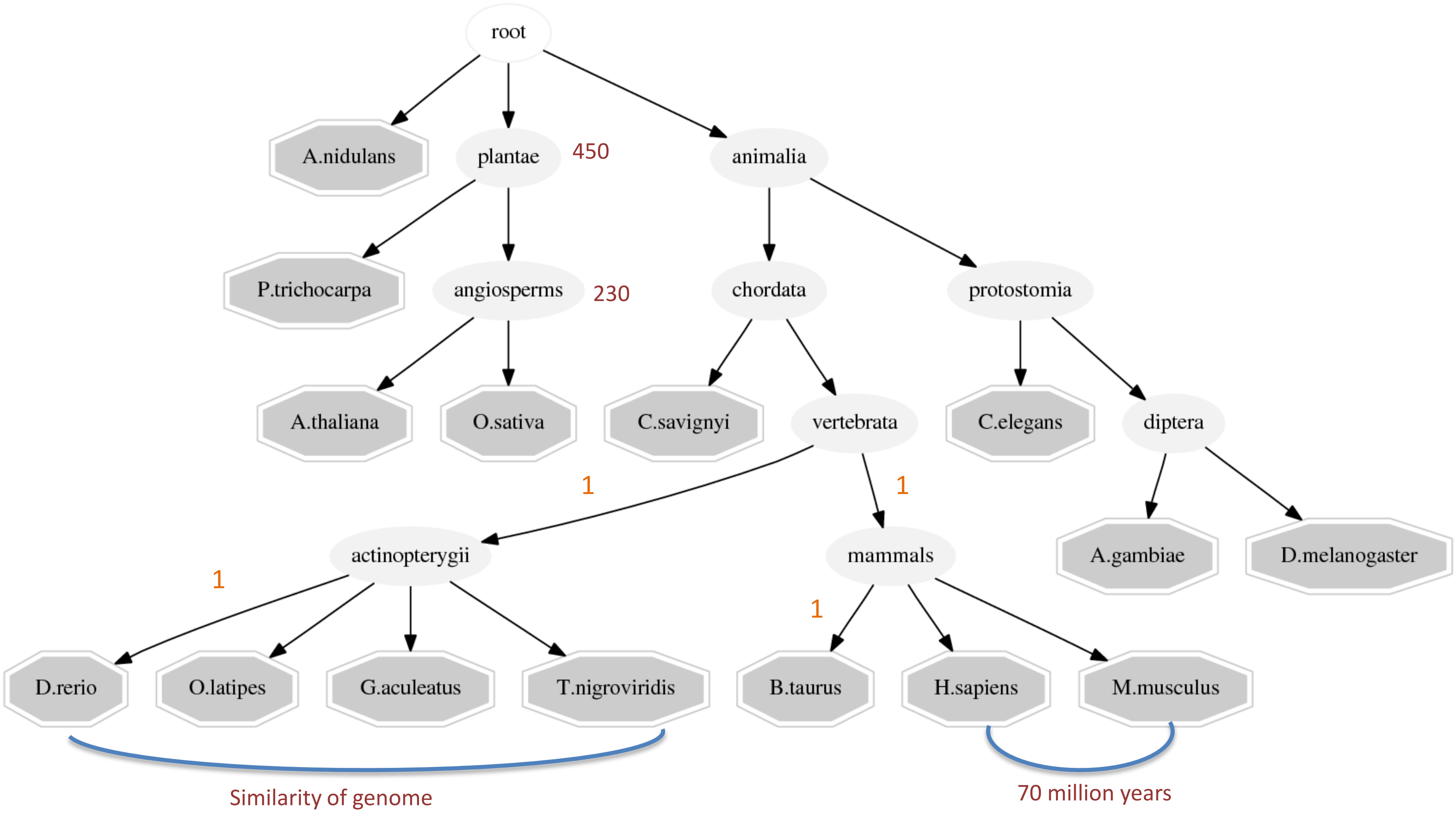


Given : annotated pre-mRNA data in multiple organisms

Find : new splicing sites ➔ find Donor and Acceptor locations

Question : Can we build a better model using data from multiple organisms?

Hierarchy used:

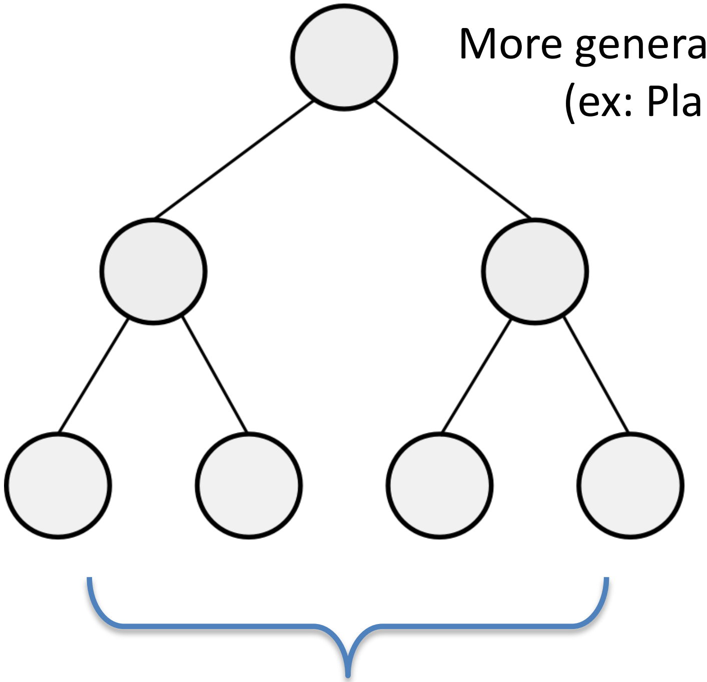


Similarity of genome

70 million years

Techniques and algorithms

Exploiting hierarchies - 1



More general organism
(ex: Plantae)

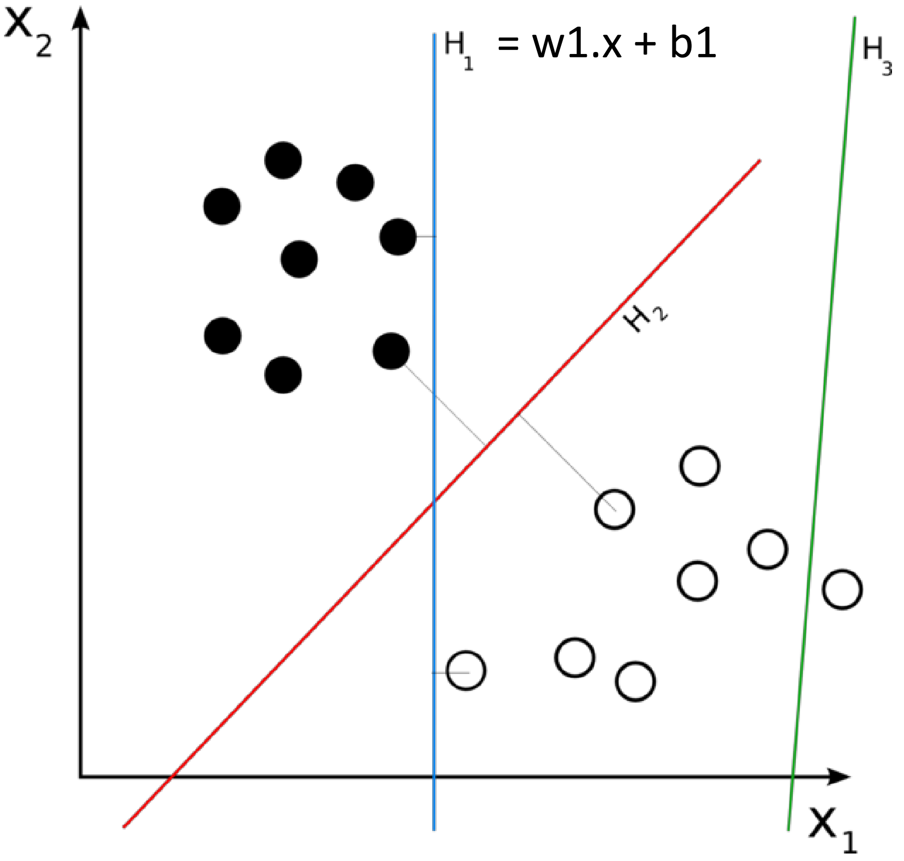
More specific organisms

Top-Down Model

1. Build model on root node using data from all leaf nodes of the tree
2. Build model on next level, similar to using all data in leaf nodes under that subtree
3. . . . . . .

- *Can use any machine learning technique to build models at each level!*
- *How to build "similar" models?*

Detour: Support Vector Machine



Mathematical formulation

$$\min_{\mathbf{w},b} \quad \underbrace{\boxed{\frac{1}{2}||\mathbf{w}||^2}}_{\text{Regularizer}} + C \sum_{(\mathbf{x},y)\in D} \underbrace{\boxed{\ell\left(\langle\Phi(\mathbf{x}),\mathbf{w}\rangle + b, y\right)}}_{\text{Error / Loss}}$$

$$\vec{\mathbf{w}} \longrightarrow \text{classifier!}$$

How to build similar models?

Given : A parent model $\vec{w}_{par}$

Want : $w \cong w_{par}$

In other words, want $(w - w_{par})$ to be small

Regular SVM

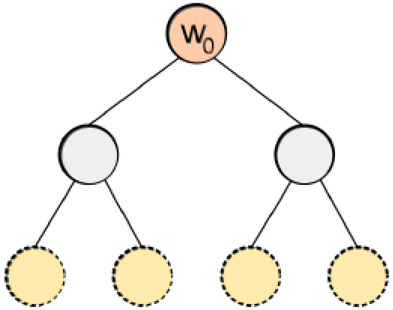
$$\min_{\mathbf{w},b} \boxed{\frac{1}{2}||\mathbf{w}||^2} + C \sum_{(\mathbf{x},y)\in D} \boxed{\ell\left(\langle \Phi(\mathbf{x}), \mathbf{w}\rangle + b, y\right)}$$

↓

DA SVM

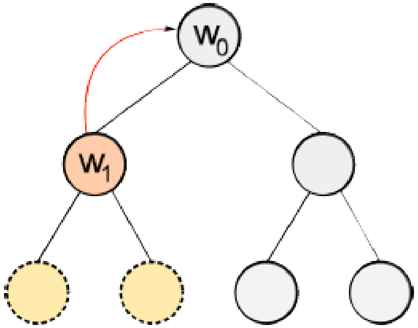
$$\min_{\mathbf{w},b} \boxed{\frac{1}{2}||\mathbf{w} - \mathbf{w}_{par}||^2} + C \sum_{(\mathbf{x},y)\in D} \boxed{\ell\left(\langle \Phi(\mathbf{x}), \mathbf{w}\rangle + b, y\right)}$$

Hierarchical Top-Down learning



(a) Top-level training

w_0 is trained using all data from all tasks (all leaf nodes)

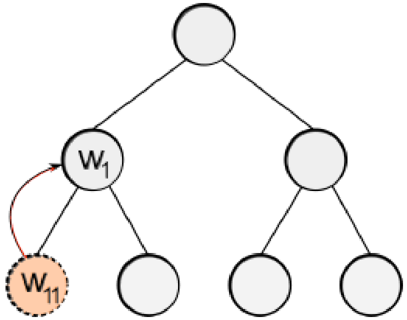


(b) Inner training

- Train on $D_i = \boxed{\bigcup_{j \preccurlyeq i} D_j}$
- Regularize $\mathbf{w}_i$ against parent predictor $\mathbf{w}_{par}$: $\|\mathbf{w}_i - \mathbf{w}_{par}\|^2$

$$\min_{\mathbf{w},b} \boxed{\frac{1}{2}\|\mathbf{w}-\mathbf{w}_{par}\|^2} + C \sum_{(\mathbf{x},y)\in D} \boxed{\ell\left(\langle\Phi(\mathbf{x}),\mathbf{w}\rangle + b, y\right)}$$

D_i

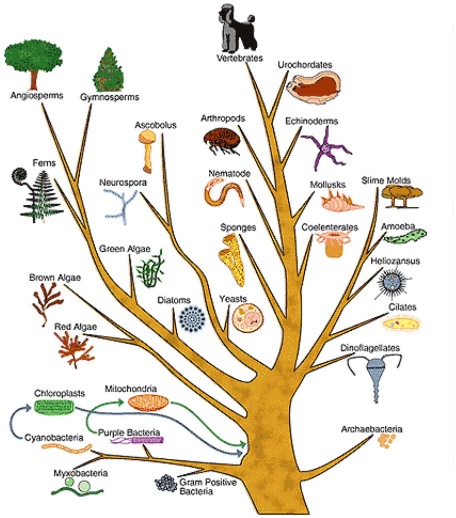


(c) Taxon training

w_{11} is built similar to w_1 and using data from one Organism

Leaf node classifiers are used for prediction

Exploiting hierarchies - 2



Task Similarity Matrix

$$\xrightarrow{\text{Transformation}} \quad \Gamma = \begin{pmatrix} \gamma_{1,1} & \cdots & \gamma_{1,M} \\ & \ddots & \\ \gamma_{M,1} & \cdots & \gamma_{M,M} \end{pmatrix}$$

↓

Pairwise &

Multitask Kernel

Pairwise Approach

- Simultaneous learning of all tasks!
- Train classifiers for all *M* tasks at the same time

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_M} \quad \frac{1}{2}\sum_{t=1}^{M}\sum_{s=1}^{M} \underset{\uparrow}{\gamma_{t,s}} \underset{\uparrow}{\boxed{\|\mathbf{w}_t - \mathbf{w}_s\|^2}} + \sum_{t=1}^{M} C_t \sum_{(\mathbf{x},y)\in D_t} \boxed{\ell\left(\langle \mathbf{x}, \mathbf{w}_t\rangle, y\right)}$$

- Similarity is enforced by the regularization term and by task similarity matrix values

Experiments and Results

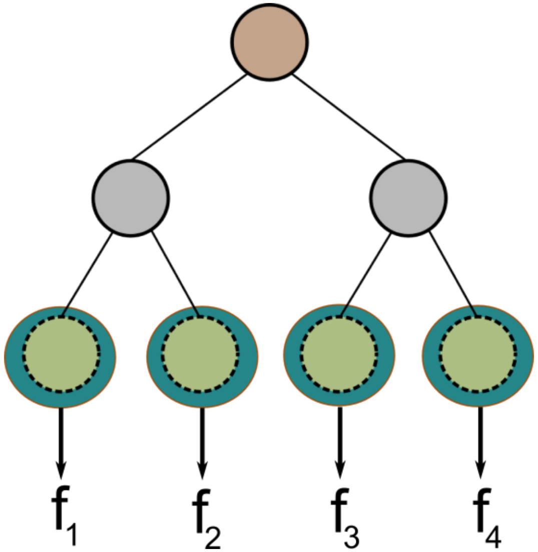
Methods compared

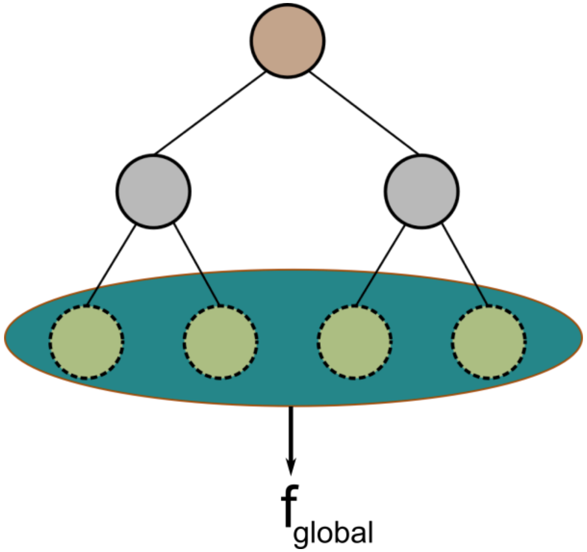
Multitask Learning Methods

1. Top-Down
2. Pairwise Regularization
3. Multitask Kernel

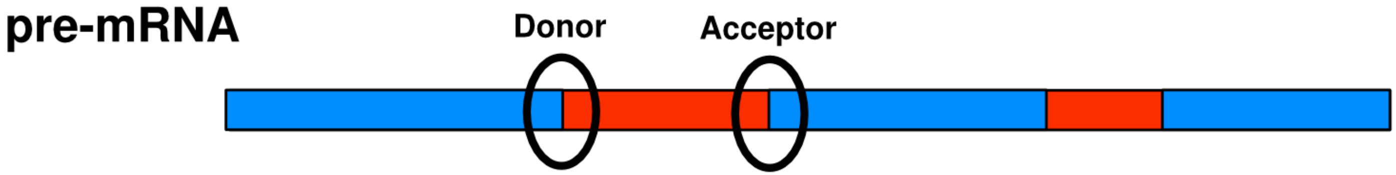
Baselines ⍰

- Plain
- Common

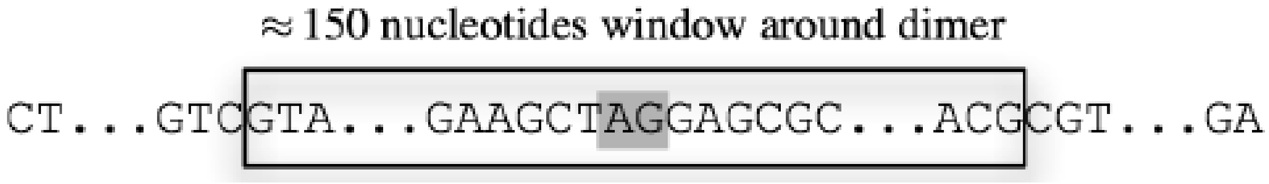




Splice-site recognition problem

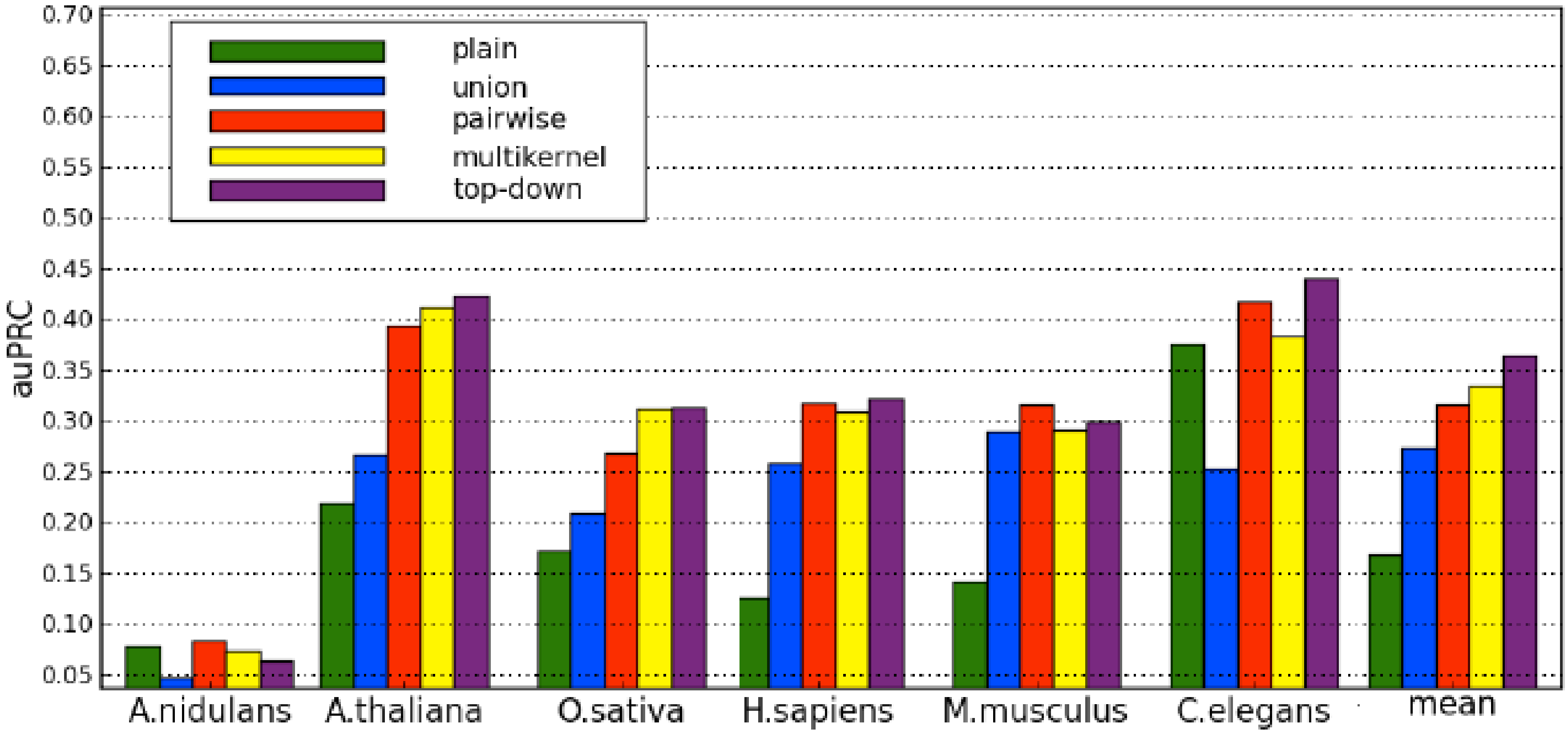


- Use insight:
  - GT or GC exhibited at donor site
  - AG consensus at acceptor site
- Each input sequence is:


≈ 150 nucleotides window around dimer
CT...GTCGTA...GAAGCTAGGAGCGC...ACGCGT...GA

- Data:
  - 15 organisms
  - Training: 10,000 examples per organism, Test data: 6,000 examples per organism

Performance : AUC
(area under precision recall curve)



Observations

- Gain is more for lower levels in hierarchy
- “A. nidulans” : baselines do better!
- “Mouse” doesn’t benefit much
    - – possible reason: not much similarity in taxonomy
- No performance loss on distantly related organisms

Critique

- Experimental evaluation not very thorough
- Learning in the absence of a hierarchy
  - How to define task similarity measures?
- Assumes that some training (labeled) data is available from all organisms
  - In many scenarios, there is no data available on less studied organisms

Comments? Thoughts?

Multitask Kernel approach

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \hat{k}(\mathbf{x}_i,\mathbf{x}_j) + \sum_{i=1}^{n}\alpha_i$$

$$\text{s.t. } 0 \le \alpha_i \le C \ \forall i \in [1,n]$$

$$\alpha^T \mathbf{y} = 0,$$

where

$$\hat{k}((x_i,s),(x_j,t)) = \underbrace{k_{\text{task}}(s,t)}_{\gamma_{t,s}} \cdot k(x_i,x_j)$$