Tree Kernel to Analyse Phylogenetic Profiles

Jean-Philippe Vert , BIOINFORMATICS, Vol. 18 Suppl.1 2002

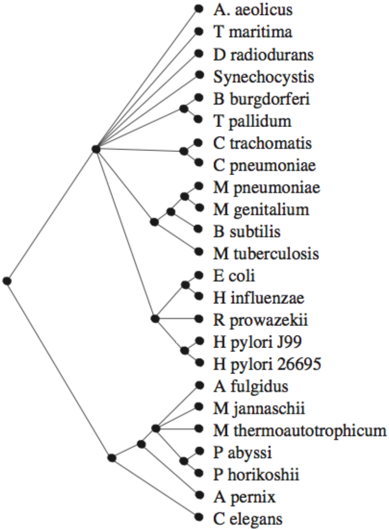
A Tree Kernel to Analyse Phylogenetic Profiles

Burçin Özcan

Department of Mathematics,
University of Houston

May 1^{st}, 2014

A **phylogenetic tree** is a tree in which each leaf of the tree represents one organism currently living and each internal node represents an ancestor of the current organisms.



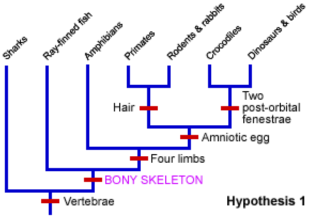
A **phylogenetic profile** is a set of bits assigned to the leaves of the phylogenetic tree.

Question : how to measure the **'similarity'** between two gene profiles to develop efficient function prediction methods

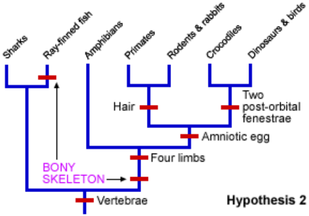
Approaches :

- counting the number of organisms where they differ (Pellegrini et. al., 1999)
- differential parsimony based on phylogenetic reconstruction of the ancestors and comparison of the trees (Liberles et. al., 2002)

Parsimony priciple is to choose the simplest explanation that fits the evidence.



Sharks
Ray-finned fish
Amphibians
Primates
Rodents & rabbits
Crocodiles
Dinosaurs & birds
Hair
Two post-orbital fenestrae
Amniotic egg
Four limbs
BONY SKELETON
Vertebrae
Hypothesis 1



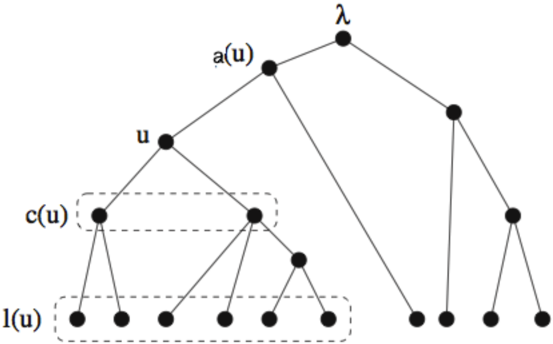
Sharks
Ray-finned fish
Amphibians
Primates
Rodents & rabbits
Crocodiles
Dinosaurs & birds
Hair
Two post-orbital fenestrae
Amniotic egg
Four limbs
BONY SKELETON
Vertebrae
Hypothesis 2

- mapping each profile to a feature space such that each coordinate corresponds to one pattern of inheritance during evolution

    1. uncertainty in ancestor tree
    2. high dimensional feature space

$$pro1, pro2 \mapsto \phi(pro1), \phi(pro2) \mapsto \langle \phi(pro1), \phi(pro2) \rangle = K_{tree}(pro1, pro2)$$

Baysesian tree models for phylogenetic profiles



- λ is the root
- T is the set of nodes of the tree
- $L \subset T$ is the nodes with no child
- $T* = T \setminus \{\lambda\}$ set of nodes without root

For any node $u \in T*$, $a(u) \in T$ is the parent of u and $c(u) \subset T$ is the set of children of u and $l(u) \subset L$ is the set of leaves descendants of u.

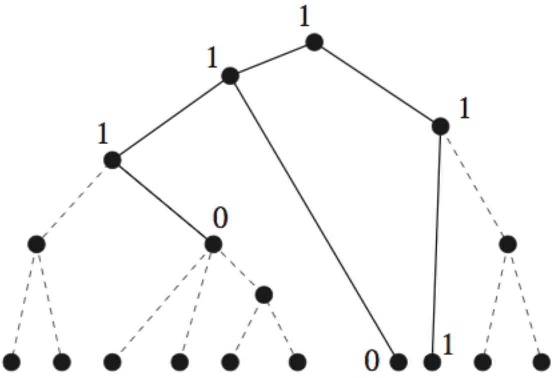


Figure : An evolution pattern

The tree is used to define a joint probability distribution P for a random set of variables $\{X_u, u \in T\}$.

$X_u = x_u \in \{0, 1\}$

To model the evolution of genomes along the branches

- an initial distribution p_λ is assigned to the root
- a conditional probability distribution $p_u(i|j)$ for (i, j) $\in \{0, 1\}^2$ is assigned to each node u $\in$ *T**.

Tree with the set of distributions $\{p_u, u \in T\}$ defines a joint probability distribution for X_T

$$p(x_T) = p_\lambda(x_\lambda) \prod_{u \in T*} p_u(x_u|x_{a(u)})$$

Kernels

- **Naive Kernel**

  $$K_{naive}(x_L, y_L) = \sum_{u \in L} x_u y_u$$

  where x_L and y_L are phylogenetic profiles.

  Note that naive kernel doesn't incorporate phylogenetic relationships among species.

- **Tree Kernel**

  An **evolution pattern** is a pair (S, z_S) where $S \subset T$ and z_S is a set of values attached to the nodes of S.

  Define the feature

  $$\phi_{S,z_S} = p(x_L|z_S)$$

  If we denote by $C(T)$ the set of all subtrees of T, the tree kernel is defined as follows:

  $$K_{tree}(x_L, y_L) = \sum_{S \in C(T)} \sum_{z_S} p(z_S)p(x_L|z_S)p(y_L|z_S)$$

Computation of the tree kernel
for each phylogenetic profile x_L using a post-order traversal of the tree,

- if u is a leaf :

$$\forall i \in \{0,1\}, \quad \mu_i(x_L, u) = p_u(x_u|i)$$

- if u is an internal node different from root:

$$\forall i \in \{0,1\}, \quad \mu_i(x_L, u) = \sum_{j \in \{0,1\}} p_u(j|i) \prod_{u' \in c(\lambda)} \mu_j(x_L, u')$$

- if u is a root:

$$\mu(x_L, \lambda) = \sum_{j \in \{0,1\}} p_\lambda(j) \prod_{u' \in c(\lambda)} \mu_j(x_L, u')$$

In order to compute the kernel between any two profiles, a second function ξ is defined

- if u is a leaf:

$$\xi_i(u) = \begin{cases} 1, & \text{if } x_u = y_u = i \\ 0, & \text{otherwise} \end{cases}$$

- if u is not a leaf :

$$\xi_i(u) = \prod_{u' \in c(u)} \left\{ \sum_{j \in \{0,1\}} p_{u'}(j|i)\xi_j(u') + \mu_i(x_L, u')\mu_i(y_L, u') \right\}$$

Tree kernel value is obtained as follows:

$$K_{tree}(x_L, y_L) = \sum_{j \in \{0,1\}} p_\lambda(j)\xi_j(\lambda) + \mu(x_L, \lambda)\mu(y_L, \lambda)$$

Data

- Experiments are performed with the genes of the budding yeast Saccharomyces cerevisiae.
- 2465 yeast genes selected.
- phylogenetic profiles are generated by computing E-value.
- parameters of Bayesian model set to $p_\lambda(1) = 0.9$, $p_u(1|1) = p_u(0|0) = 0.9$
- function prediction experiments based on CYGD containing several hundred functional classes, but used only the classes with at least 10 genes, resulting in 133 classes.

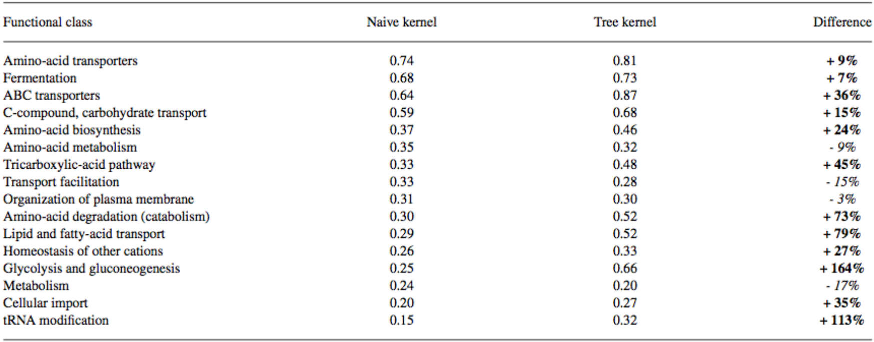
To infer the function of a gene from its phylogenetic profile, SVM is used.

To get a glimpse of the relative positions of the phylogenetic profiles in the feature space, kernel PCA analysis is performed.

Receiver Operating Characteristic (ROC curve) is a graphical plot which illustrates the performance of a binary classifier as its discrimination threshold is varied. It is created by plotting the TPR vs. FPR.

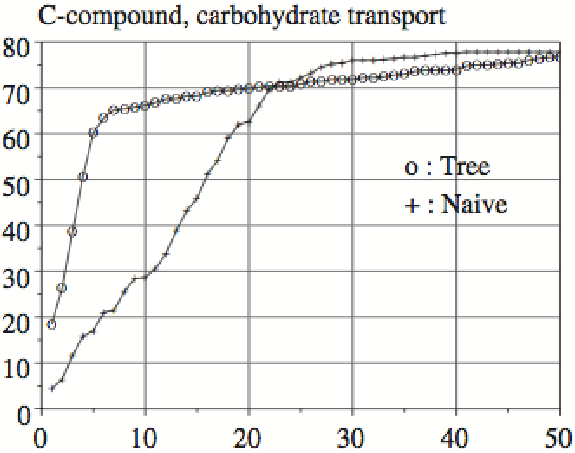
ROC_{50} score is the area under the ROC curve plotted until 50 true negatives are found. The number of negatives is fixed at 50.

Results

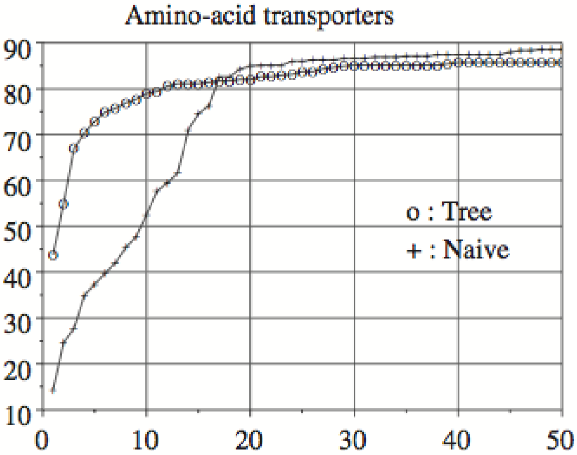


Functional class	Naive kernel	Tree kernel	Difference
Amino-acid transporters	0.74	0.81	**+ 9%**
Fermentation	0.68	0.73	**+ 7%**
ABC transporters	0.64	0.87	**+ 36%**
C-compound, carbohydrate transport	0.59	0.68	**+ 15%**
Amino-acid biosynthesis	0.37	0.46	**+ 24%**
Amino-acid metabolism	0.35	0.32	*- 9%*
Tricarboxylic-acid pathway	0.33	0.48	**+ 45%**
Transport facilitation	0.33	0.28	*- 15%*
Organization of plasma membrane	0.31	0.30	*- 3%*
Amino-acid degradation (catabolism)	0.30	0.52	**+ 73%**
Lipid and fatty-acid transport	0.29	0.52	**+ 79%**
Homeostasis of other cations	0.26	0.33	**+ 27%**
Glycolysis and gluconeogenesis	0.25	0.66	**+ 164%**
Metabolism	0.24	0.20	*- 17%*
Cellular import	0.20	0.27	**+ 35%**
tRNA modification	0.15	0.32	**+ 113%**

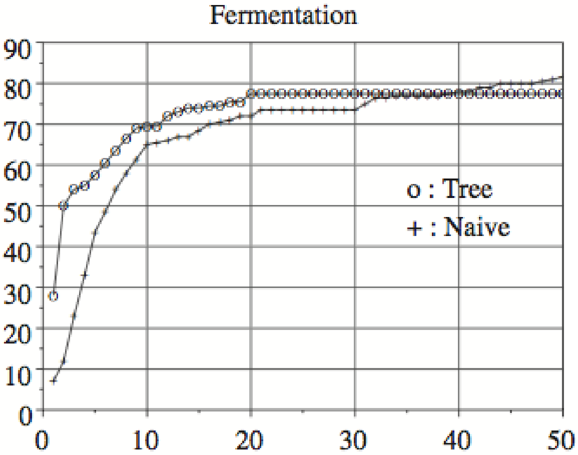
Figure : ROC_{50} scores for the prediction of 16 functional categories by a SVM


C-compound, carbohydrate transport
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
o : Tree
+ : Naive

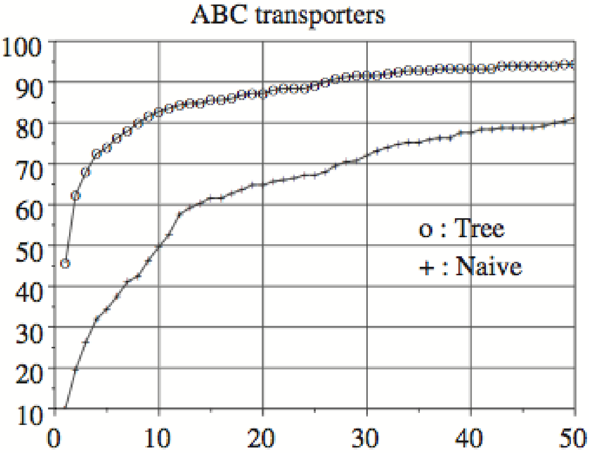
ROC curve for prediction of functional class C-compound from the phylogenetic profiles of the yeast genes with SVM using tree and naive kernel.
The number of false positives is on the X-axis, the percentage of true positives on the Y-axis.


Amino-acid transporters
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
o : Tree
+ : Naive

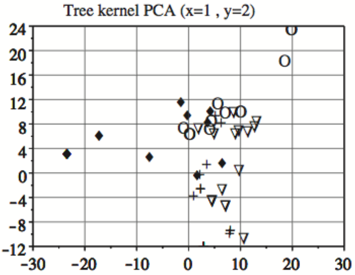
ROC curve for prediction of the functional class Amino-acid transporters from the phylogenetic profiles of the yeast genes with SVM using tree and naive kernel.
The number of false positives is on the X-axis, the percentage of true positives on the Y-axis.

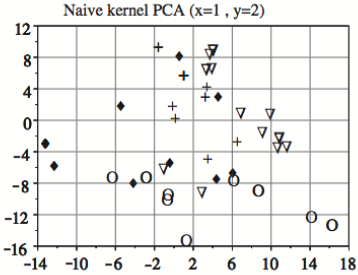

Fermentation
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
o : Tree
+ : Naive

ROC curve for prediction of the functional class Fermentation from the phylogenetic profiles of the yeast genes with SVM using tree and naive kernel.
The number of false positives is on the X-axis, the percentage of true positives on the Y-axis.


ABC transporters
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
o : Tree
+ : Naive

ROC curve for prediction of the functional class ABC transporters from the phylogenetic profiles of the yeast genes with SVM using tree and naive kernel. The number of false positives is on the X-axis, the percentage of true positives on the Y-axis.

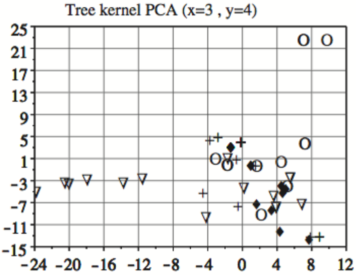

Tree kernel PCA (x=1 , y=2)
24
20
16
12
8
4
0
-4
-8
-12
-30
-20
-10
0
10
20
30

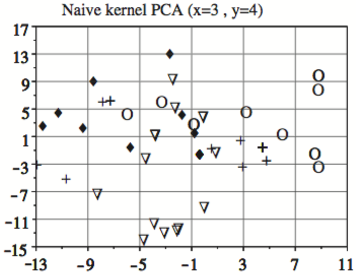

Naive kernel PCA (x=1 , y=2)
12
8
4
0
-4
-8
-12
-16
-14
-10
-6
-2
2
6
10
14
18

Kernel PCA analysis of the phylogenetic profiles of the genes belonging to 4 functional classes, with tree kernel on the left and naive kernel on the right. the first PC vs the second PC.

The profiles are more scattered in all dimensions in feature space with navie kernel than the tree kernel.

With the naive kernel, the profiles seem to to form a convex set, but with tree kernel, each PC seems to characterize particularly a few profiles, which happen to belong to the same functional class. For instance, Amino-acid transporters for the first component.


Tree kernel PCA (x=3 , y=4)


Naive kernel PCA (x=3 , y=4)

Kernel PCA analysis of the phylogenetic profiles of the genes belonging to 4 functional classes, with tree kernel on the left and naive kernel on the right. the third PC vs the forth PC.

The profiles are more scattered in all dimensions in feature space with navie kernel than the tree kernel.

With the naive kernel, the profiles seem to to form a convex set, but with tree kernel, each PC seems to characterize particularly a few profiles, which happen to belong to the same functional class. For instance, ABC transporters for the third component.

Conclusion

- the geometry defined by the tree kernel is more sensitive to biologically relevant patterns than when no evolutionary information is used
- SVM using the tree kernel performed on average better than a SVM using a naive kernel

Thank you