A tree kernel to analyze phylogenetic profiles

Jean-Philippe Vert

Bioinformatics Center, Kyoto University, Japan
Jean-Philippe.Vert@mines.org

10th International Conference on Intelligent Systems for Molecular Biology
(ISMB02), August 3-7, 2002, Edmonton, Canada

Outline

1. Phylogenetic profiles

2. The tree kernel

3. Implementation

4. Experimental results

Part 1

Phylogenetic profiles

Definition

- The phylogenetic profile of a gene is a vector of bits which indicates the presence (1) or absence (0) of orthologs in every fully sequenced genome.

Gene	aero	aful	. . .	tpal	worm
YAL001C	1	1	. . .	0	0
YAB002W	0	0	. . .	0	1
⋮	⋮	⋮	⋮	⋮	⋮

- Can be estimated *in silico* by sequence similarity search

From profile to function

- Genes are likely to be transmitted together in evolution when they participate:
  - ⋆ to a common structural complex,
  - ⋆ to a common pathway.
- Consequently genes with similar phylogenetic profiles are likely to have similar functions
- How to measure the similarity between profiles?

Naive approach

- Count the number of bits in common:

x	1	1	0	1	0	0	0	1	1	0
y	1	0	1	0	0	0	0	1	0	1

$$s(x,y) = 5$$

- Cluster or use k-NN for gene function prediction with this similarity measure (Pellegrini et al., 1999)

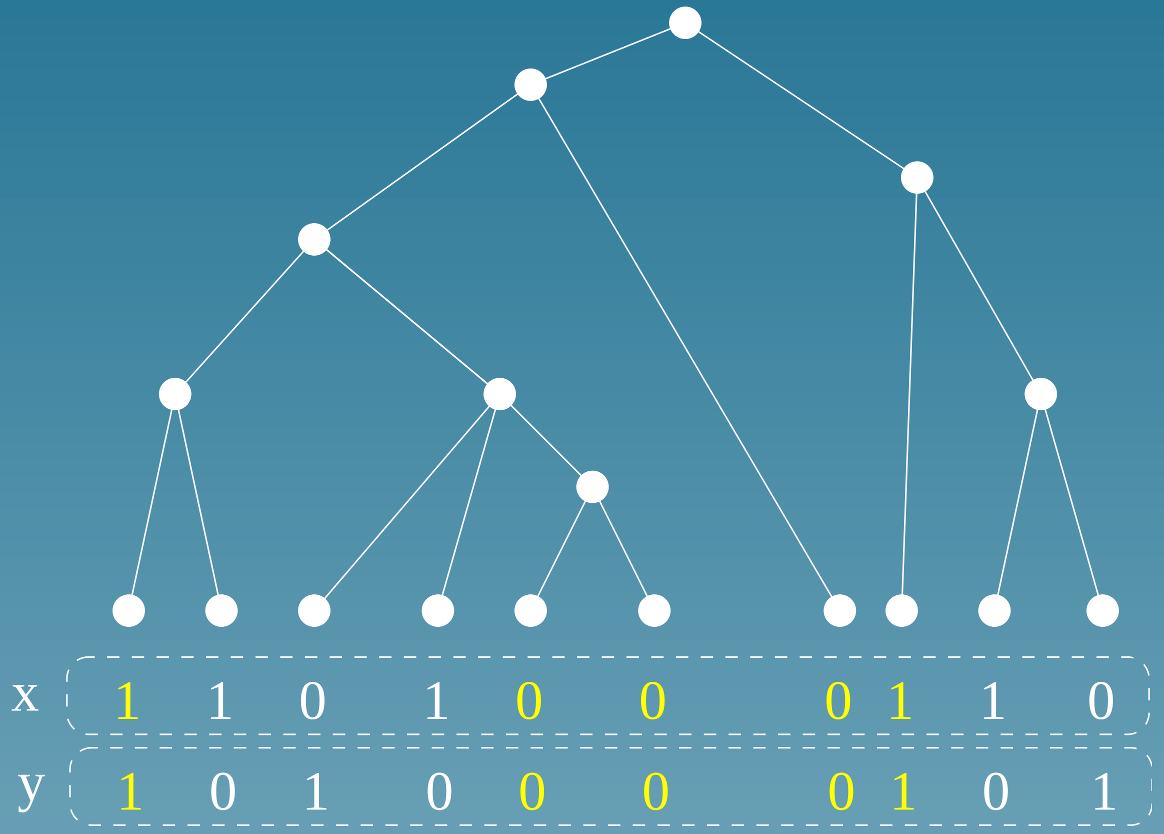
Limitations of the naive approach

- The set of sequenced organisms has a strong influence on the similarity score (e.g., eukaryotes are under-represented)

- A more detailed understanding of when two proteins were transmitted together or not during evolution could be useful

- A function could be characterized by only a subset of the bits (e.g., 1 in eukaryotes, 0 in bacteria, whatever in archae)

What is not used in the naive approach



x 1 1 0 1 0 0 0 1 1 0

y 1 0 1 0 0 0 0 1 0 1

The knowledge of the phylogenetic tree that links the species together.

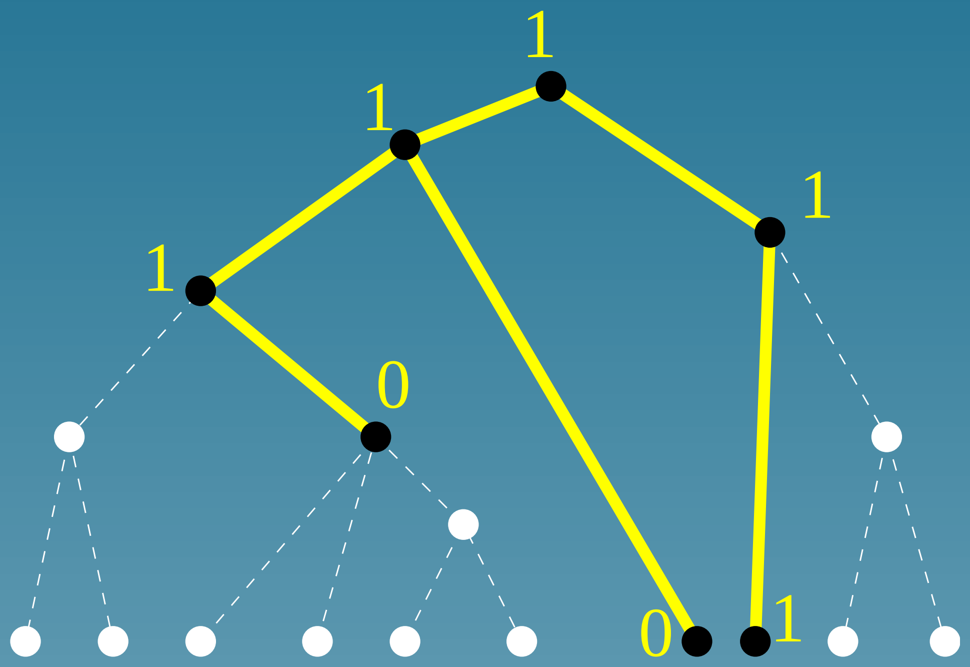
Part 2

The tree kernel

Overview

- Profiles are mapped to a high-dimensional vector space (*feature space*).
- Each coordinate in the feature space is an evolutionary relevant pattern (e.g.,*the gene was transmitted in eukaryotes and bacteria, but not in archae*)
- It is possible to work implicitly in the feature space and to use powerful classification algorithms (support vector machines).

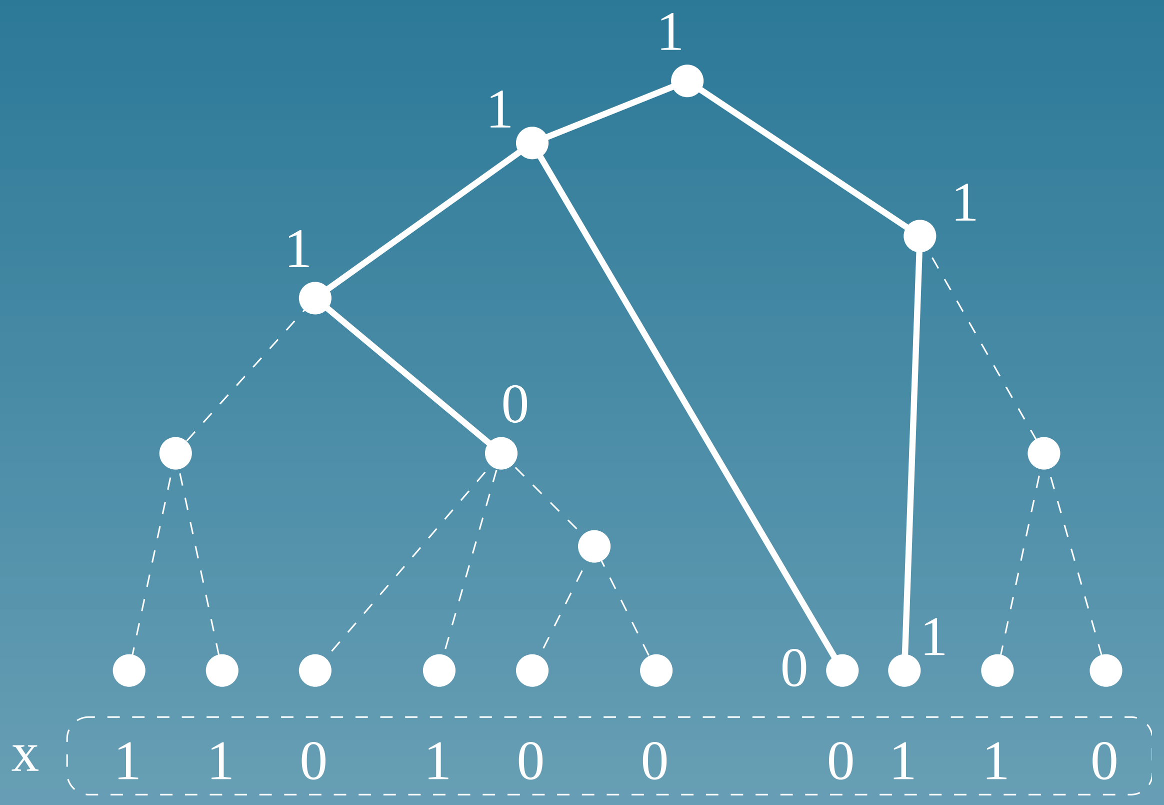
Evolution pattern



- A possible pattern of transmission during evolution

- Mathematically, a rooted subtree with nodes labeled 0 or 1.

Evolution patterns and phylogenetic profiles



Impossible to know for sure if the gene followed exactly this evolution pattern

Probabilistic model of gene transmission

- The phylogenetic tree as a tree graphical model

- Simplified model:

  ⋆ $P(1) = 1 - P(0) = 0.9$, at the root,

  ⋆ Along each branch transmission follows the transition matrix:

$$\begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$

Probabilistic assignment of evolution pattern

For a phylogenetic profile x and an evolution pattern:

- $P(e)$ quantifies how "natural" the pattern is
- $P(x|e)$ quantifies how likely the pattern e is the "true history" of the profile x

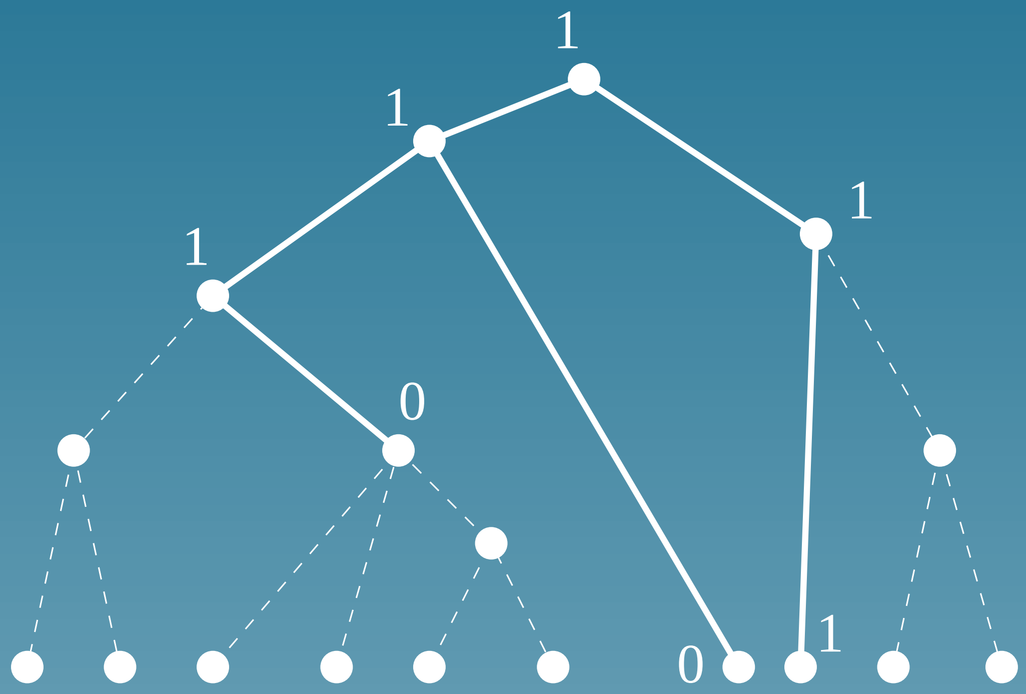
Representation of a profile in terms of evolution patterns

- Consider all possible evolution patterns $(e_1, \ldots, e_N)$. A profile x can be represented by the N-dimensional vector:

$$\Phi(x) = \begin{pmatrix} \sqrt{P(e_1)}P(x|e_1) \\ \vdots \\ \sqrt{P(e_N)}P(x|e_N) \end{pmatrix}$$

- Comparing $\Phi(x)$ and $\Phi(y)$ gives a precise idea of which evolution patterns are shared or not by x and y.

Comparing two profiles through evolution patterns



x	1	1	0	1	0	0	0	1	1	0
y	1	0	1	0	0	0	0	1	0	1

Tree kernel

- Kernel methods (SVM, kernel-PCA, kernel-clustering...) only require the computation of the kernel function:

$$K(x, y) = \Phi(x).\Phi(y).$$

- In our case we obtain the tree kernel:

$$K(x, y) = \sum_{e} P(e)P(x|e)P(y|e),$$

where the sum is over all possible evolution patterns.

Part 3

Implementation

The problem

- For any two profiles x and y we need to compute:

$$K(x,y) = \sum_{e} P(e)P(x|e)P(y|e).$$

- For kernel methods such as SVM, the computation of the kernel should be as quick as possible (limiting factor)

- The number of expression patterns in the sum increases exponentially with the length of the profiles...

Trick 1

- For any given pattern e, the term:

$$\alpha(e) = P(e)P(x|e)P(y|e)$$

  can be factorized and computed recursively by working up the tree from the leaves

- Classical trick for computing likelihood with graphical models, cf. Felsenstein's algorithm

Trick 2

- The sum
$$\sum_{e} \alpha(e)$$
over all subtrees can also be factorized and computed recursively by working up the tree from the leaves

- Similar in spirit to the Context Tree Weighting algorithm (Willems et al., 1995).

Combining tricks

- Both tricks can be combined (see proceedings)
- $K(x, y)$ can be computed by two post-order traversals of the tree
- The complexity is linear with the length of the profile.

Part 4

Experimental results

Gene function prediction with SVM

- Profiles for 2465 genes of *S. Cerevisiae* were computed by BLAST search (cf Pavlidis et al. 2001), using 24 genomes.

- Consensus phylogenetic tree (cf. Liberles et al. 2002) with simplified probabilistic model of gene transmission

- SVM trained to predict all functional classes of the MIPS catalog with at least 10 genes (cross-validation)

- Comparison of the tree kernel with the naive kernel

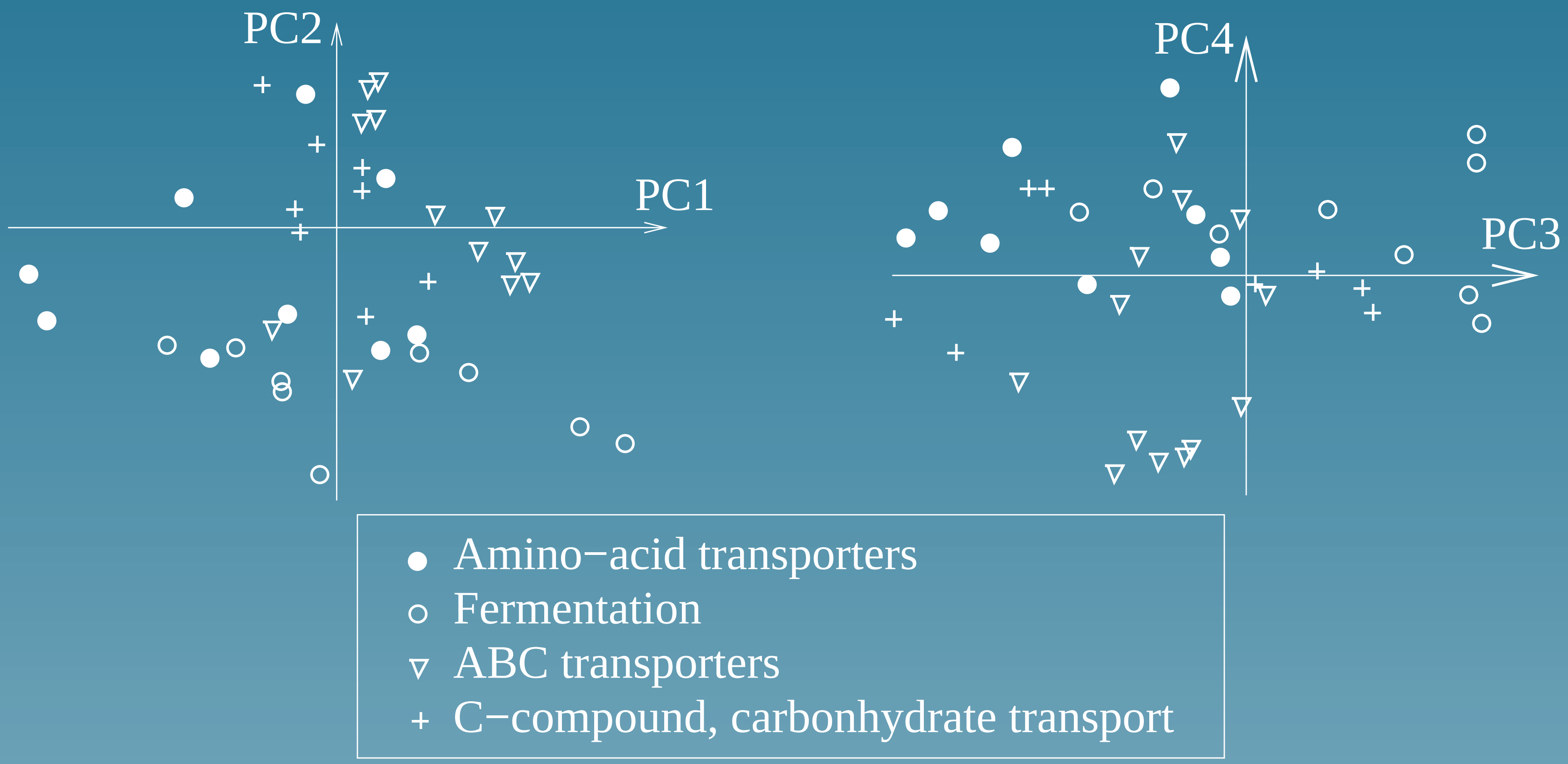
Results (ROC 50)

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 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%*

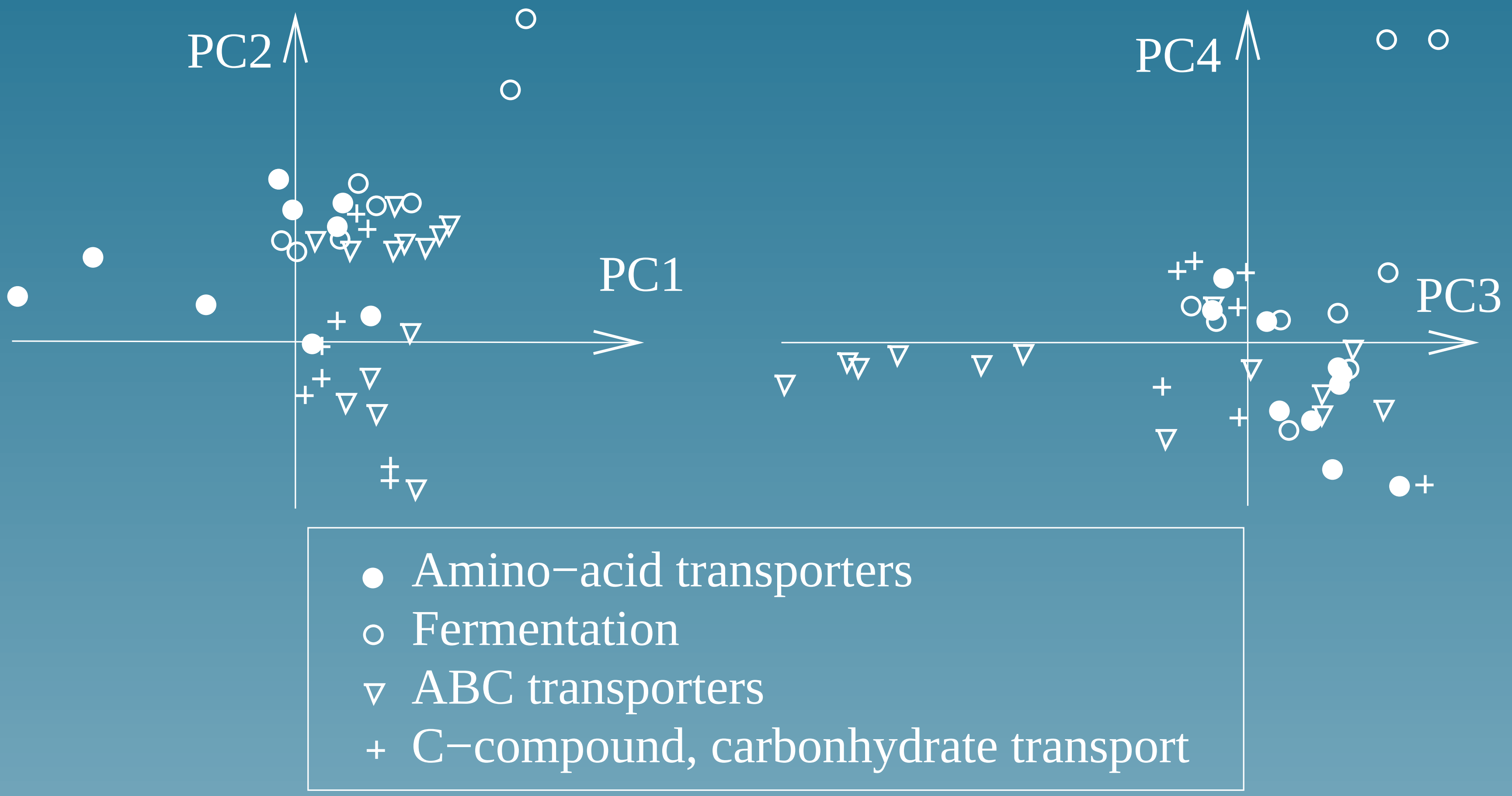
A insight into the feature space

- PCA can be performed implicitly in the feature space with a kernel function: kernel-PCA (Scholkopf et al. 1999)
- Projecting the genes on the first principal components gives an idea of the shape of the features space

Naive kernel PCA



Tree kernel PCA



Conclusion

Conclusion

- The tree kernel $K(x, y)$ is a similarity measure for phylogenetic profiles
- Two profiles are similar is they are likely to have shared many evolution patterns
- $K(x, y)$ can be efficiently computed
- $K(x, y)$ can be used by any kernel method
- Phylogenetic profiles are not only vectors of bits.