A combinatorial test for significant codivergence between cool-season grasses and their symbiotic fungal endophytes

Ruriko Yoshida
Dept. of Statistics University of Kentucky

Joint work with C.L. Schardl, S. Speakman, K.D. Craven, A. Lindstrom, A. Stromberg

Endophytes species

Transmission strategies

Relative importance of vertical vs horizontal transmission.

A. Exclusively or almost exclusively transmitted horizontally. The endophyte will tend to shut down host seed production ("choke disease"), diverting available plant resources to production of infectious spores. Those spores spread to developing seeds of neighboring plants.

B. Exclusive vertical transmission. The host exhibits no disease symptoms due to the endophyte infection. Its seeds develop and germinate normally, but bear the endophyte and thereby transmit it to the next generation.

C. Mixed vertical and horizontal transmission strategy.

Host range

1. Some endophytes are restricted to individual host species. This seems rare for endophyte categories A and C above, but typical of category B.

2. Some are restricted to individual genera.

3. Some are restricted to host tribes.

4. Some are associated mainly with one host tribe, but occasionally can be identified in the sister tribe.

5. Some are present in a phylogenetically broad range of host tribes.

Problem

Question. We would like to analyze how grasses and their endophytes evolved together?

Method.

1. We use phylogenetic trees among grass species and among endophytes species.

2. Compute pairwise distances in the grass tree and in the endophyte tree.

3. Compute **MRCA pairs** of two trees.

4. Estimate the probability of codivergence between two trees and compute their correlations.

Phylogenetic trees

Data. Sequencing of Chloroplast DNA (cpDNA) Non-Coding Regions. 27 species in each group. Sequences were entered into GenBank as accession numbers AY450932–AY450949 and EU119353–EU119377.

Based on published phylogenetic inference for the grass subfamily Poöideae (Soreng and Davis 1998), *Brachyelytrum erectum* was chosen as the outgroup for reconstructing the grass phylogenies. The corresponding endophyte, *Epichloë brachyelytri*, was the outgroup chosen for endophyte phylogenies.

We reconstruct phylogenetic trees of grasses (the host tree, T_H) and phylogenetic trees of endophytes (the parasite tree, T_P) via a software `PAUP*` under the GTR+G+I model.

Then we used a software `r8s` to make trees unltrametric (using the least square method).

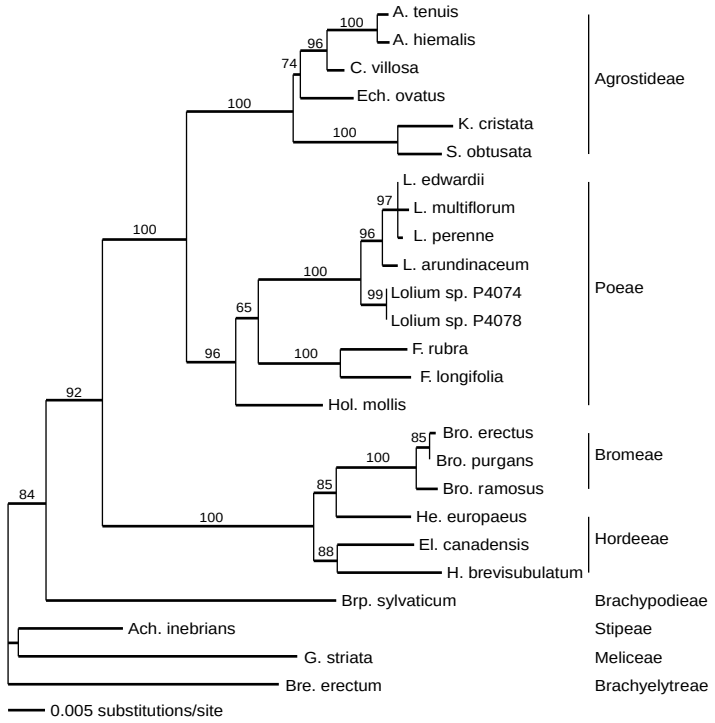


Figure 1: Parametric ML tree estimated from cpDNA intron and intergenic sequences. Numbers above branches indicate bootstrap support percentages (over 50%) obtained by 1000 maximum parsimony searches with branch swapping.

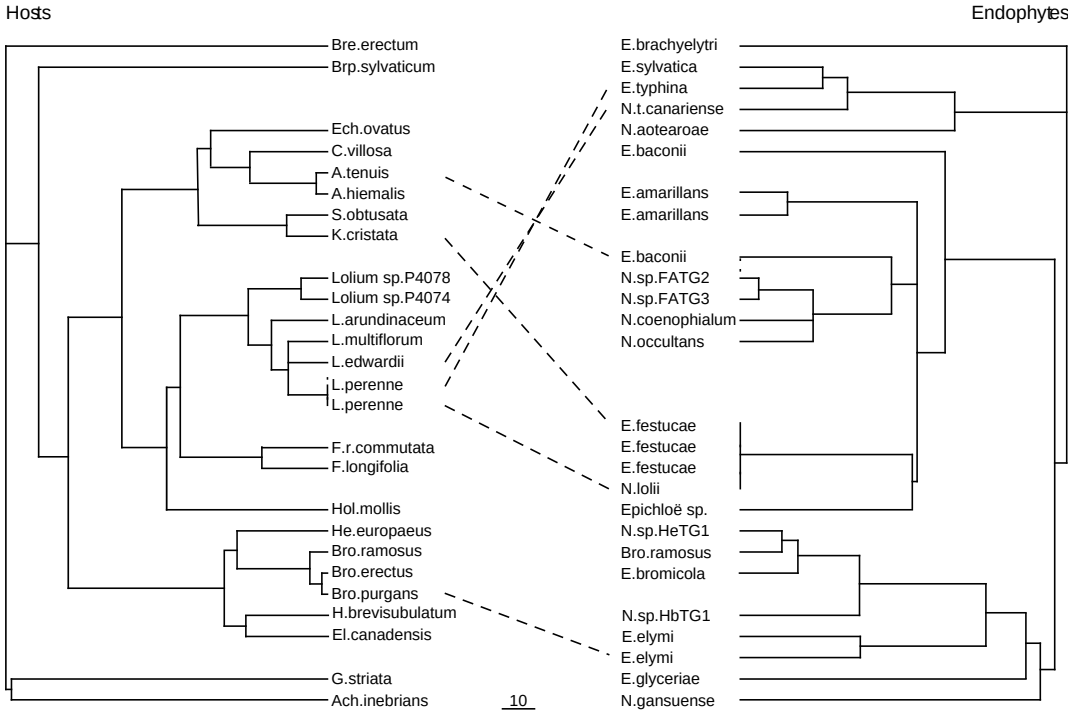


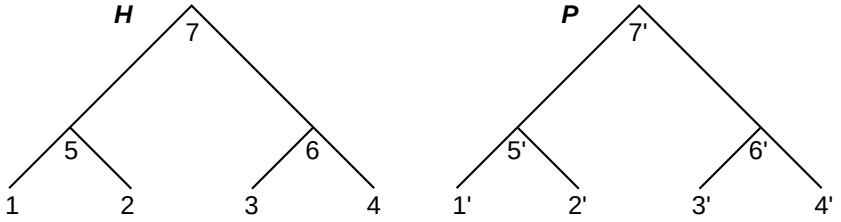
Figure 2: Ultrametric ML time trees for host grasses and their endophytes. Hosts and their endophytes are indicated opposite each other or by connecting dashed lines. Full taxon names are given in Table 1 in our paper.

MRCA pairs

A **MRCA pair** is a pair of a Most Recent Common Ancestor (MRCA) of any pair of host species and a Most Recent Common Ancestor (MRCA) of any pair of parasite species.

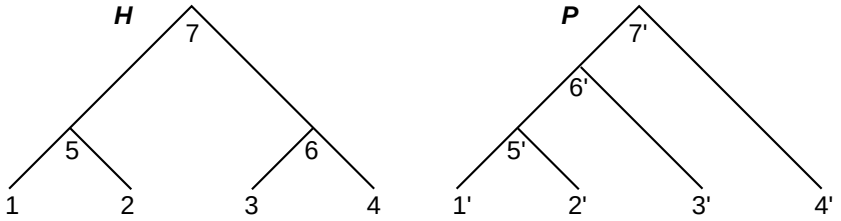
MRCA pairs

Congruent trees



MRCA pair	Pairs of *H* and *P* taxon pairs
(5,5')	((1,2),(1',2'))
(6,6')	((3,4),(3',4'))
(7,7')	((1,3),(1',3')), ((1,4),(1',4')), ((2,3),(2',3')), ((2,4),(2',4'))

Incongruent trees



MRCA pair	Pairs of *H* and *P* taxon pairs
(5,5')	((1,2),(1',2'))
(7,6')	((1,3),(1',3')), ((2,3),(2',3'))
(7,7')	((1,4),(1',4')), ((2,4),(2',4'))
(6,7')	((3,4),(3',4'))

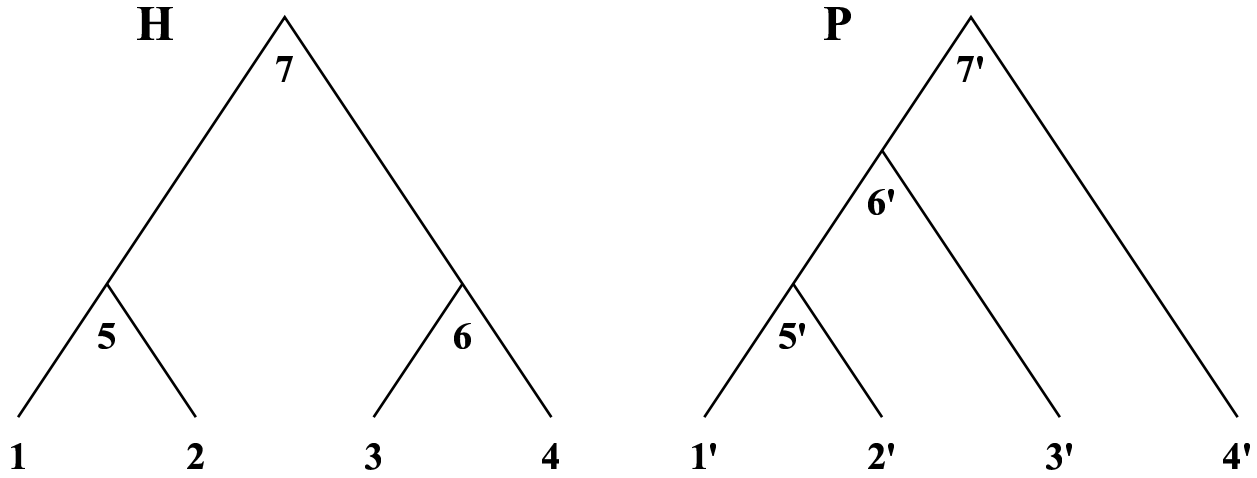
Analysis on codivergence

[Legendre et al 2002] etc used all possible pairs of pairwise distances from the host tree and the parasite tree and used Principal Components Analysis (PCA) to compute their correlations.

A problem of their method is that we possibly pick the same Most Recent Common Ancestor (MRCA) pair multiple times. This causes a bias in the result. In each tip clade a MRCA uniquely relates two taxa. However, a MRCA deeper in the tree relates multiple taxon pairs. For example, for congruent H and P trees the matrix of all pairwise distances of H taxon pairs against all pairwise distances of P taxon pairs represents each corresponding pair of tip clade MRCAs only once, and each corresponding pair of deeper MRCAs multiple times. The **MRCALink algorithm** samples corresponding H and P MRCA pairs only once.

MRCALink algorithm

We will go though the algorithm with an example.



Step 1: Assign each node a unique number such that its number is bigger than its children.

Step 2: **for** each interior node in H, from all possible pairs of offsprings, find corresponding pairs in P.

5: From $5 = (1, 2)$, we find a new MRCA $5' = (1', 2')$ in P.

6: From $6 = (3, 4)$ we find a new MRCA $7' = (3', 4')$.

7: From $7 = (1, 3) = (2, 3) = (2, 4)$, we find new MRCAs $6' = (1', 3')$ and $7' = (1', 4')$.

Thus, we have pairs $(5, 5')$, $(6, 7')$, $(7, 7')$, $(7, 6')$.

Computing the probability of codivergence

Let τ_H be the set of all ultrametric host trees with n taxa and let τ_P be the set of all ultrametric parasite trees with n taxa.

$$S(X, Y, T, t) = \sum_{x \in X, y \in Y} |\text{time}(\text{MRCA}(x)) - \text{time}(\text{MRCA}(y))|,$$

where $T \in \tau_H$, $t \in \tau_P$, X is a set of pairs of taxa in H, and Y is a set of pairs of taxa in P.

Then we estimate the probability

$$P(S(X, Y, T_H, T_P) \leq S(X, Y, T, t) : \forall T \in \tau_H, \forall t \in \tau_P)$$

which is the estimated probability of codivergence for T_H and T_P, by randomly generated trees from τ_H and τ_P.

Results

We analyzed 4 pairs of host trees and parasite trees, namely the full tree and T_1–T_4 by removing some of species in the full trees, trimmed trees (T_1–T_4).

For each pair of trimmed trees, we removed some species from the endophytes and corresponding grasses because these endophytes seem to have holizontal or mixed transmission.

Table 1: The p-values obtained by applying the dissimilarity method to all pairwise distances (noted by ALL) and to the MRCALink-derived matrix (noted by MRCA) for full and T_1 – T_4 plant and endophyte data sets (see Table 1 for the data sets). SF means a sampling fraction.

Method	Data	SF = 0.0005	SF = 0.001	SF = 0.5	SF = 1.0
ALL	Full	0.784	0.783	0.677	0.374
MRCA	Full	0.123	0.123	0.081	0.039
ALL	T_1	0.117	0.115	0.035	0.009
MRCA	T_1	< 0.001	< 0.001	< 0.001	< 0.001
ALL	T_2	0.093	0.085	0.027	0.012
MRCA	T_2	< 0.001	< 0.001	< 0.001	< 0.001
ALL	T_3	0.064	0.061	0.017	0.005
MRCA	T_3	< 0.001	< 0.001	< 0.001	< 0.001
ALL	T_4	0.018	0.020	0.005	0.002
MRCA	T_4	< 0.001	< 0.001	< 0.001	< 0.001

Cophylogeny

Cophylogeny

Suppose we have two sets of multi-species sequence data H and P. A common task in phylogenetics is to infer a tree T_H for H, or T_P for P.

Let $\mathcal{T}_H$ be the space of trees on H and $\mathcal{T}_P$ be the space of trees on P.

A **cophylogeny** is a pair of trees $(T_H, T_P) \in \mathcal{T}_H \times \mathcal{T}_P$. Usually in a cophylogeny, the trees T_H and T_P are related.

Example: H is a set of host species, and P is a set of corresponding parasite species. Or, H is a set of species, and P is a set of corresponding orthologous genes in the species.

Statistical/machine learning methods for cophylogeny

Distances between trees

We are applying distances on tree structures to assess codivergences in related trees (such trees might be for hosts and parasites (or symbionts), or they may be for distinct, putatively orthologous genes in genomes).

In order to use tools such as linear classifiers, we need biologically meaningful inner products on trees.

Why we care?

If we find some outlier trees, then they might represent noncanonical evolutionary processes such as

- Horizontal transfer of genes between species.
- Ancient polymorphisms maintained by balancing selection.
- Paralogs that may be difficult to distinguish from orthologs by other means.
- Radically different evolutionary rates between genes.

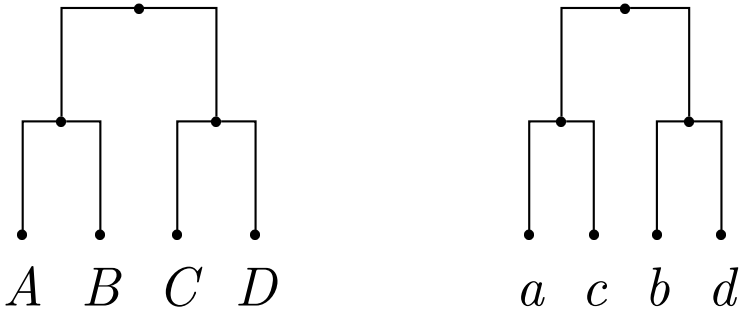
If we find multiple clusters, then they might represent recombinations.

Inner products on trees

We are particularly interested in distances $d(T_1, T_2)$ on trees which can be expressed by an inner product K in some vector space representation, i.e. $d(T_1, T_2) = \sqrt{\{K(T_1 - T_2, T_1 - T_2)\}}$. Examples include

- the l_2 inner product on $\mathbb{R}_+^{\binom{n}{2}}$, the space of dissimilarity maps
- the l_2 inner product on $\mathbb{R}_+^{\binom{n}{2}}$, the space of edge matrices of trees (k-interval).
- The l_2 inner product on $\mathbb{R}_+^{3\cdot\binom{n}{4}}$, the space of quartets whose i th element is 1 if the tree T has the particular quartet and is 0 if not.

Example: k-intervals



Recall: a k-interval distance between these trees is 2 and the difference between each pair of leaves can be written as a matrix:

$$\begin{pmatrix} 0 & 2 & 4 & 4 \\ 2 & 0 & 4 & 4 \\ 4 & 4 & 0 & 2 \\ 4 & 4 & 2 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 4 & 2 & 4 \\ 4 & 0 & 4 & 2 \\ 2 & 4 & 0 & 4 \\ 4 & 2 & 4 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 2 & 2 & 0 \\ 2 & 0 & 0 & 2 \\ 2 & 0 & 0 & 2 \\ 0 & 2 & 2 & 0 \end{pmatrix}$$

The distance is the l_2 norm of this vector.

Comparing distributions instead of point estimates of trees

Given host/parasite (or genes) sequence data H and P, respectively, a standard method for comparing host/parasite trees is to compute a fixed host tree $\hat{T}_H$ and parasite tree $\hat{T}_P$, and then compute $d(\hat{T}_H, \hat{T}_P)$, and then take $d(\hat{T}_H, \hat{T}_P)$ to be the true distance between host and parasite (or gene) trees.

But there is uncertainty in host/parasite trees, and point estimates of trees can be unreliable. Given distributions $\mathbb{D}_H$ and $\mathbb{D}_P$ on host and parasite trees (conditional on host/parasite sequence data), we could instead compare the distributions.

Example: Difference-of-means testing

Given distributions $\mathbb{D}_H$ and $\mathbb{D}_P$, a classical quantity of interest in statistics is the difference of means: $d(\mathbb{E}_{\mathbb{D}_H} T_H, \mathbb{E}_{\mathbb{D}_P} T_P)$.

We can perform difference of means testing for host and parasite tree distributions

Suppose we have distance $d(T_H, T_P)$ defined between trees, given by an inner product

$$d(T_H, T_P) = \sqrt{K(T_H - T_P, T_H - T_P)}$$

in some feature space.

If we define $d(\mathbb{D}_H, \mathbb{D}_P) := d(\mathbb{E}_{\mathbb{D}_H} T_H, \mathbb{E}_{\mathbb{D}_P} T_P)$, we obtain a metric on tree distributions. Note this can be written entirely in terms of the inner product K:

$$d(\mathbb{D}_H, \mathbb{D}_P)^2 :=$$

$$-2\mathbb{E}_{\mathbb{D}_H \times \mathbb{D}_P} K(T_H, T_P) + \mathbb{E}_{\mathbb{D}_H \times \mathbb{D}_H} K(T_H^{\{1\}}, T_H^{\{2\}}) + \mathbb{E}_{\mathbb{D}_P \times \mathbb{D}_P} K(T_P^{\{1\}}, T_P^{\{2\}})$$

Upshot: If we have an oracle to compute $K()$, then we can estimate $d(\mathbb{D}_H, \mathbb{D}_P)^2$ via MCMC without writing down vector representations of trees or means. This is an example of a kernel method in machine learning.

Important when vector space is high dimensional but inner product can be computed quickly. For example if feature vectors are quartet indicators, then dimension is $O(n^4)$, but inner product can be computed in $O(n \log n)$ time.

We would like to be able to determine whether $d(\mathbb{D}_H, \mathbb{D}_P)$ is significantly greater than zero.

Now we have the statistical hypothesis:

$$H_0 : d(\mathbb{D}_H, \mathbb{D}_P)^2 = 0$$

$$H_1 : d(\mathbb{D}_H, \mathbb{D}_P)^2 > 0$$

Bootstrap

We can bootstrap columns of H to obtain bootstrapped sets of hypothetical host data $\hat{H}$, and similarly bootstrap P to obtain sets of hypothetical parasite data $\hat{P}$.

Then we determine whether $d(\mathbb{D}_H, \mathbb{D}_P)$ looks significantly large by counting the number of bootstraps satisfying

$$d(\mathbb{D}_H, \mathbb{D}_P) < d(\mathbb{D}_H, \mathbb{D}_{\hat{H}}) + d(\mathbb{D}_P, \mathbb{D}_{\hat{P}}) \text{ for each bootstrap } \hat{H}, \hat{P}.$$

The p-value for our statistical test is the frequency of the counting.

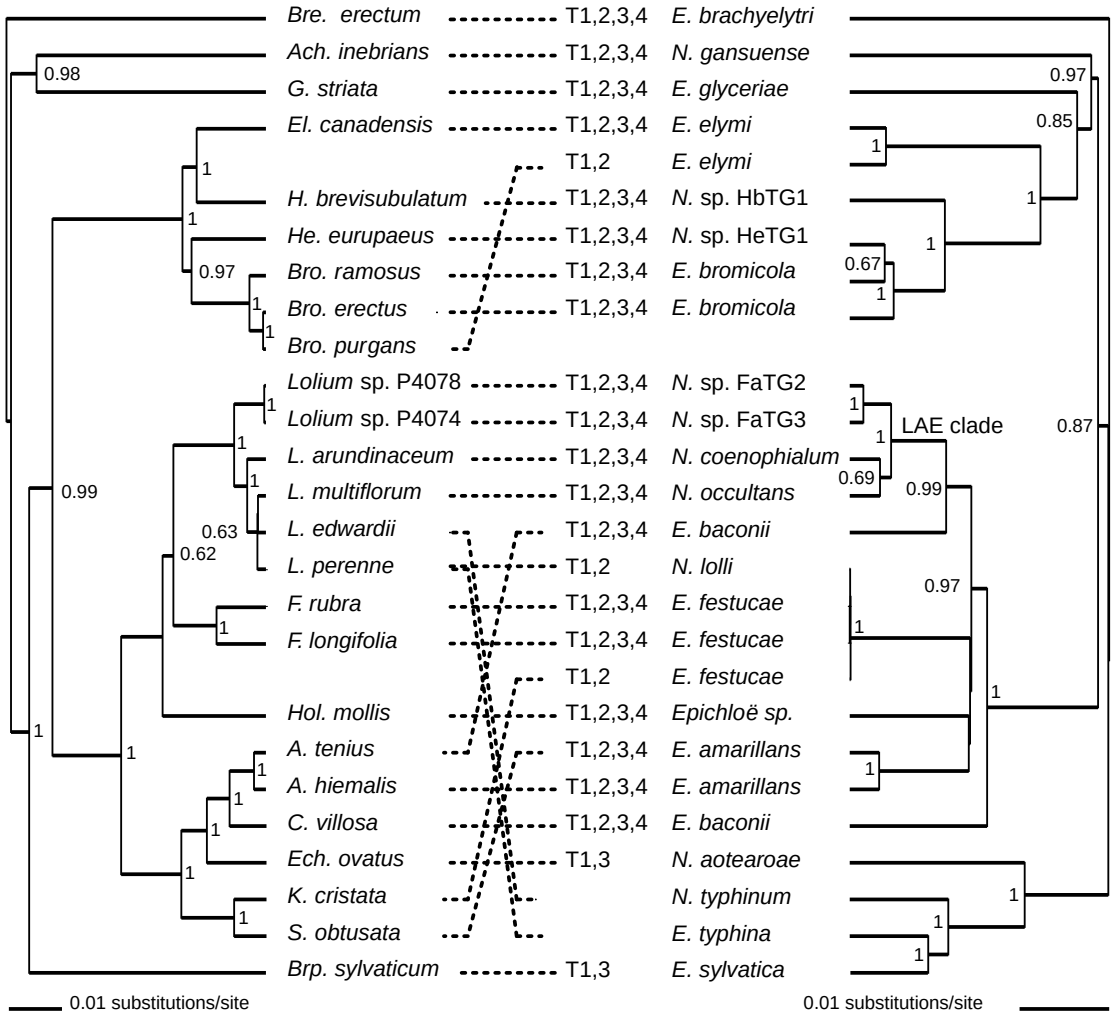


Figure 3: Ultrametric ML time trees for plant and endophyte data sets in [Schardl et al, 2008] constructed via BEAST. Hosts and their endophytes are indicated by dashed lines. Numeric values on nodes represent their posterior probabilities estimated by BEAST.

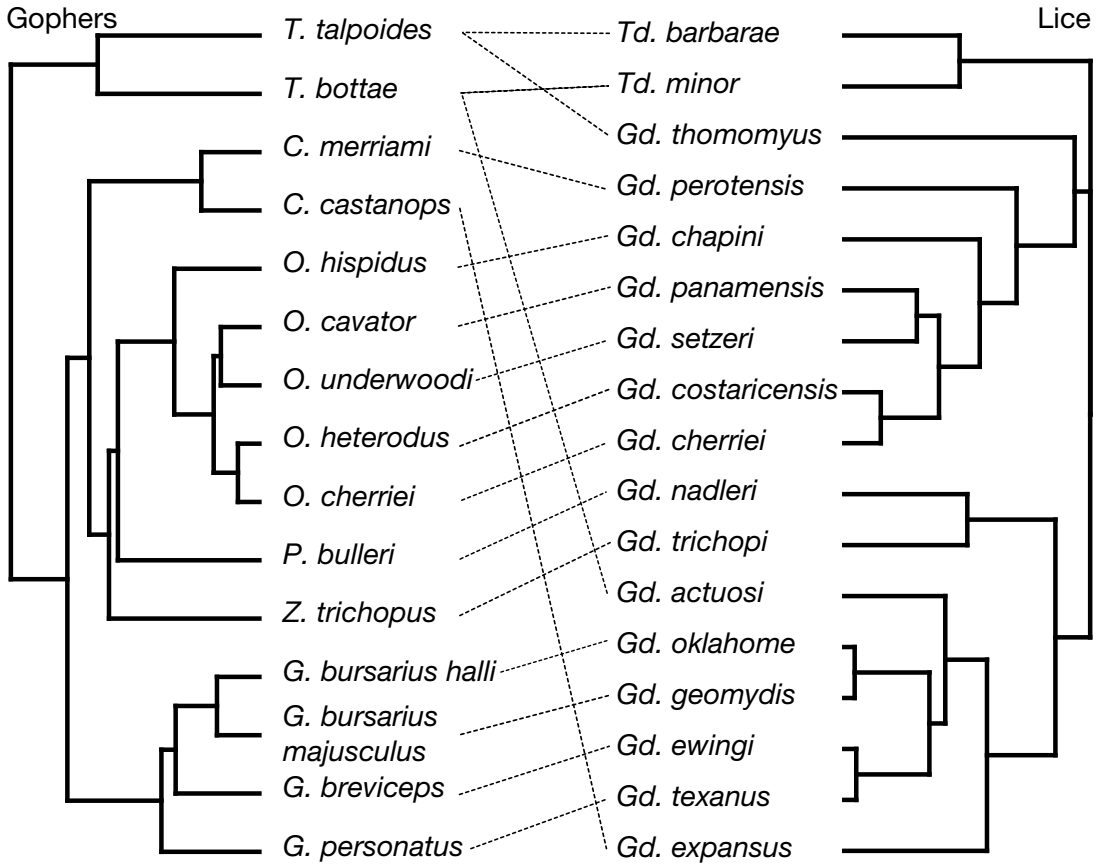


Figure 4: Ultrametric ML time trees for gopher and louse data sets [Hafner, 1990] constructed via `BEAST`. [Page and Hafner, 1994] and [Huelsenbeck et. al., 2000] studied these data sets.

Results

$60,000$ sampled tree via MCMC and 100 bootstrap with k-interval kernel.

For plant-endophytes data sets we got the p-value < 0.001.

For gopher and louse data sets we got the p-value $= 0.14$.

Applications to the fungal housekeeping genes: From Kerry O'Donnell, with the *Epichloë festucae* genome: gene ATUB, BTUB, EF1alpha, HIS, ITS, MAT, PHO84, RED, TRI101, TRI3, URA.

The TRI3 and TRI101 genes are reported to conflict with each other, even though they are in the same gene cluster and involved in the same process: synthesis of toxic trichothecenes.

Our test shows that the p-value $= 0.02$. Also we found some small p-value (0.20) between ATUB and ITS, but we think it is because the ITS tends to be badly homoplastic.

Other kernel methods for comparing tree distributions

- Rather than test for difference in means, we can also find a plane which gives the "best separation" between host and parasite tree distributions.

- Method: MCMC sample a cloud of host trees, and a cloud of parasite trees, and then find the best separating hyperplane (linear decision boundary) between the clouds, in some vector space.

- In machine learning, SVMs can be used for this task. SVMs can be run as a kernel method: decision boundary is expressed in terms of host and parasite trees, without ever writing down explicit vector representations.

- Splitting hyperplane tells us *how* the host and parasite tree distributions are different: what features (e.g. which pairs of taxa, or which quartets) give the highest disagreement between host and parasite tree distributions.

Future work

SVMs for tree distributions

Can we define similar statistical/machine learning methods, using geodesic distance measure?

Are there other more on the space of cophylogenetic trees which are "biologically meaningful"?

Advertisement

The Sepcial Session on Advances in Algebraic Statistics

Organized by Sonja Petrović and RY

2010 AMS Spring Southeastern Sectional Meeting

Lexington, KY, March 27–28, 2010 (Saturday – Sunday)

`http://www.ams.org/amsmtgs/2162_program_ss2.html#title`

Reference

in Systematic Biology. Volume 57, Issue 3, (2008), p483 – 498

Available at `http://arxiv.org/abs/q-bio.PE/0611084`

Thank you....