Geometry of Cophylogeny

Ruriko Yoshida
Dept. of Statistics University of Kentucky

Joint work with M. Owen and P. Huggins

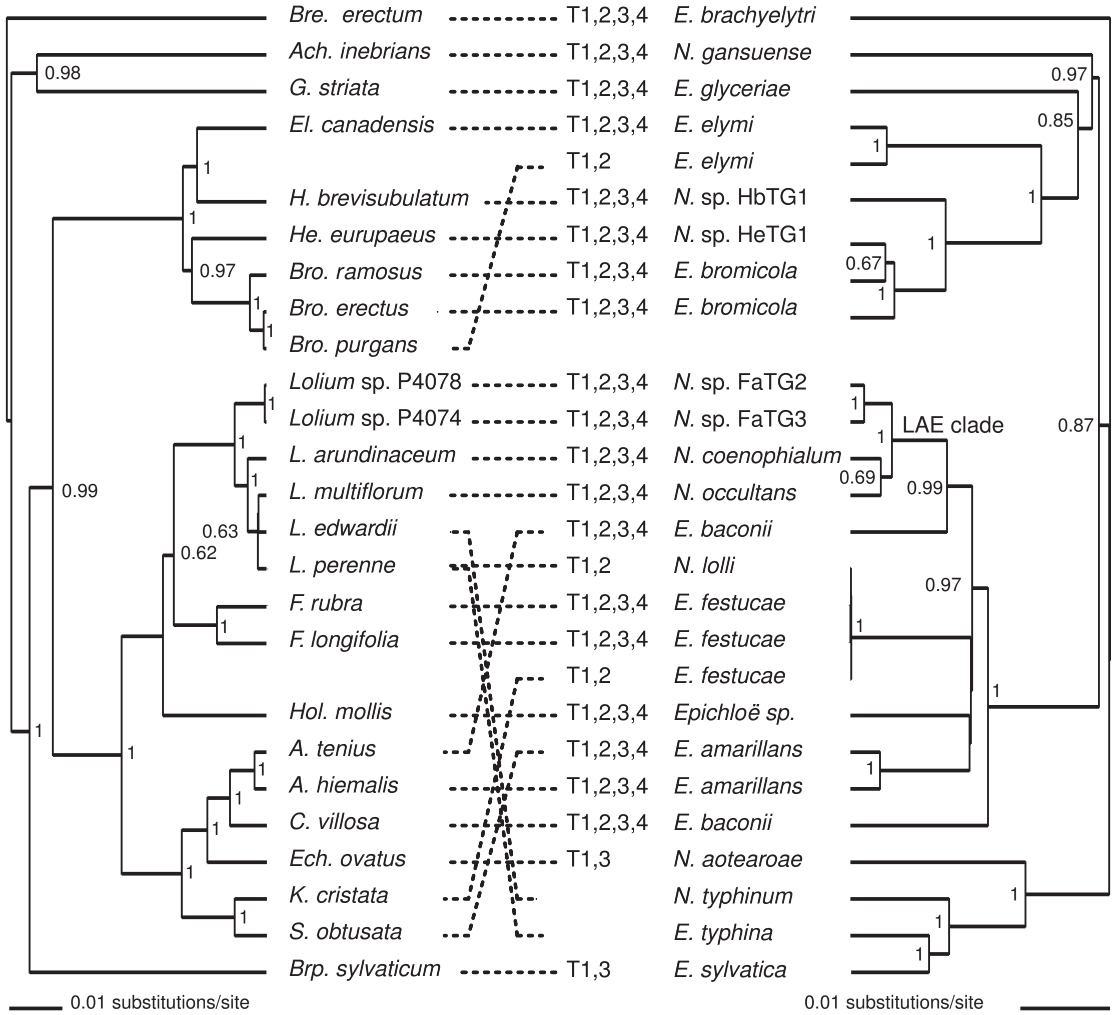


Figure 1: Ultrametric ML time trees for plant and endophyte data sets in [Schardl et al, 2008] constructed via BEAST. Hosts and their endophytes are indicated by dashed lines. Numeric values on nodes represent their posterior probabilities estimated by BEAST.

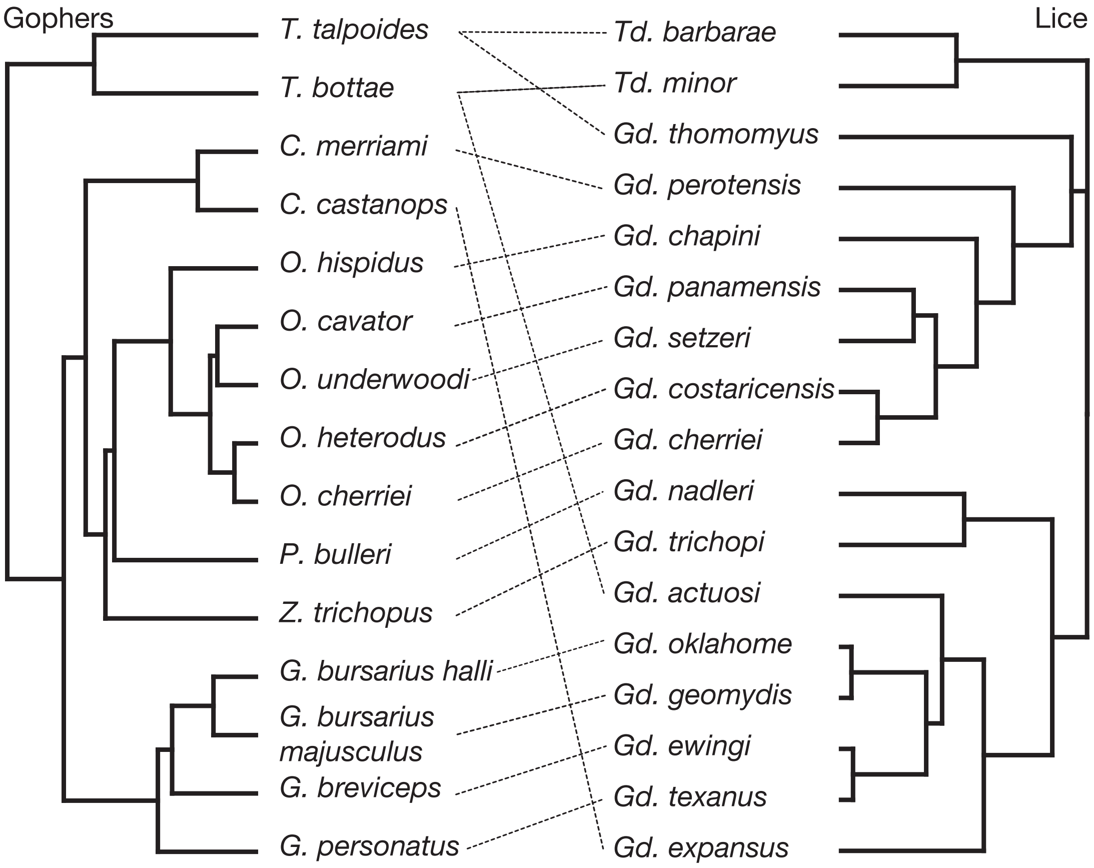


Figure 2: Ultrametric ML time trees for gopher and louse data sets [Hafner, 1990] constructed via BEAST. [Page and Hafner, 1994] and [Huelsenbeck et. al., 2000] studied these data sets.

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.

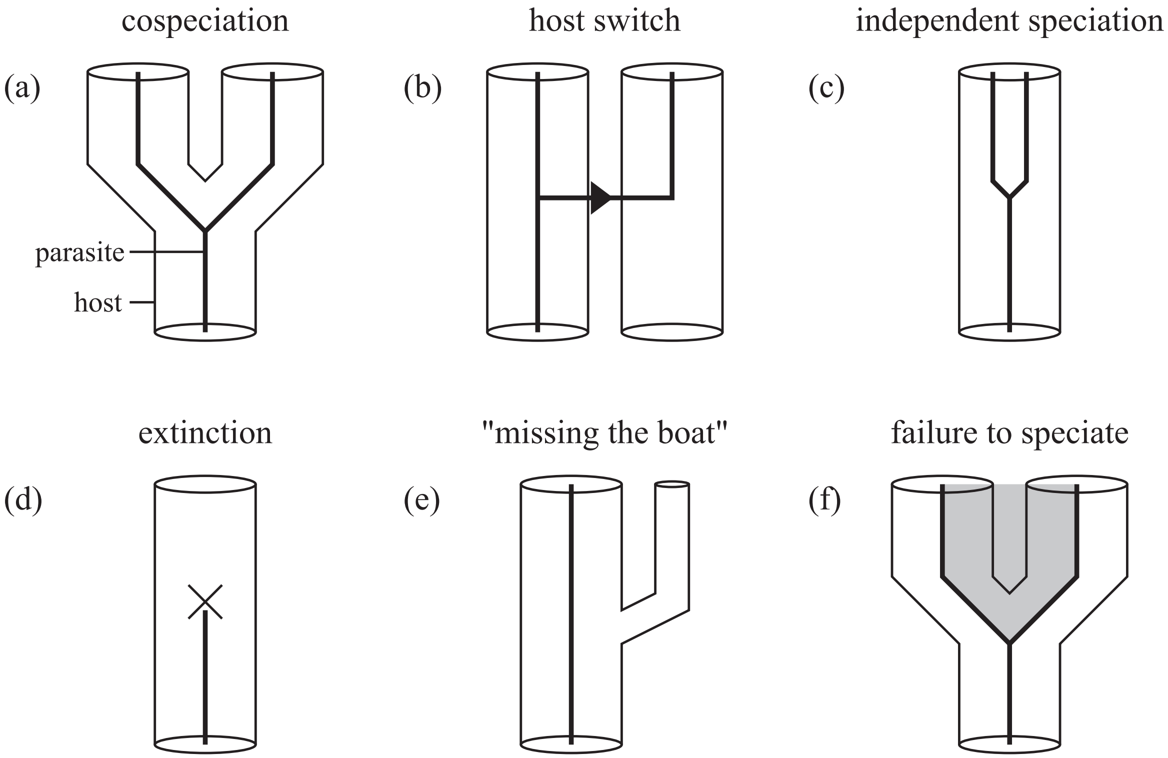
A **cophylogeny** is a pair of trees $(T_H, T_P) \in \mathcal{T}_H \times \mathcal{T}_P$. Usually in a cophylogeny, the trees T_H and T_P are related.

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.

Want: Propose a research program of extending ideas of phylogeny to cophylogeny.

6 different processes in a host–parasite association

Even though two phylogenetic trees are coevolved, tree topologies of T_H and T_P might differ.



Geometry of Cophylogenetic trees

Definition: A subset $S \subset \mathcal{T}_H \times \mathcal{T}_P$ is called a **space of cophylogenetic trees**.

Definition: The projection $S_{T_H} = \{T_P : (T_H, T_P) \in S\} \subset \mathcal{T}_P$ is called the **space of cophylogenetic trees given T_H**.

Example: If we assume a perfect codivergence, that is, T_H and T_P are identical (for e.g., [Huelsenbeck et. al., 2000]), the space of cophylogenetic trees is the diagonal $\{T_H = T_P\}$ of $\mathcal{T}_H \times \mathcal{T}_P$

More examples of cophylogenies...

- Coalescent cophylogeny: T_H is a species tree, and T_P is a gene tree generated from T_H according to the coalescent process.

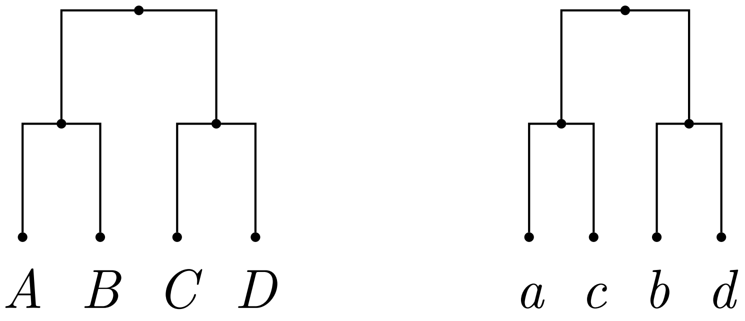
- $\leq \epsilon$-distance cophylogeny: If T_H, T_P are in the space of cophylogenetic trees, then $d(T_H, T_P) \leq \epsilon$, where d is a metric on tree space.

- Example metrics on tree space:
  - $d(T_H, T_P) :=$ geodesic length between T_P and T_H in tree space. (Studied by M. Owen, also developed software for geodesic length)
  - $d(T_H, T_P) :=$ quartet distance
  - $d(T_H, T_P) :=$ SPR distance
  - $d(T_H, T_P) :=$ NNI distance
  - $d(T_H, T_P) :=$ R-F symmetric difference

The space of k-interval cospeciation

In host-parasite evolution, a speciation in host is likely to be followed by a reactionary speciation in parasite, and often vice versa. Combinatorially, this assumption can be made explicit by assuming that for each pair of host species A, B, and corresponding parasite species a, b, the number of edges between A, B is within k of the number of edges between a, b. We say such a cophylogeny satisfies **k-interval cospeciation**.

Example



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 space of 1-interval cospeciation

Proposition [Huggins, Owen, and Y., 2008]

Under the 1-interval cospeciation with the given host tree T_H in taxa $\{1, 2, \cdots, n\}$, if a tree T_P in taxa $\{1', 2', \cdots, n'\}$ contains a quartet $[i'_1, i'_3; i'_2, i'_4]$ or $[i'_1, i'_4; i'_2, i'_3]$, and if the corresponding quartet in T_H generated by their hosts $\{i_1, i_2, i_3, i_4\}$ is $[i_1, i_2; i_3, i_4]$, then T_P cannot be the parasite tree for T_H.

Example

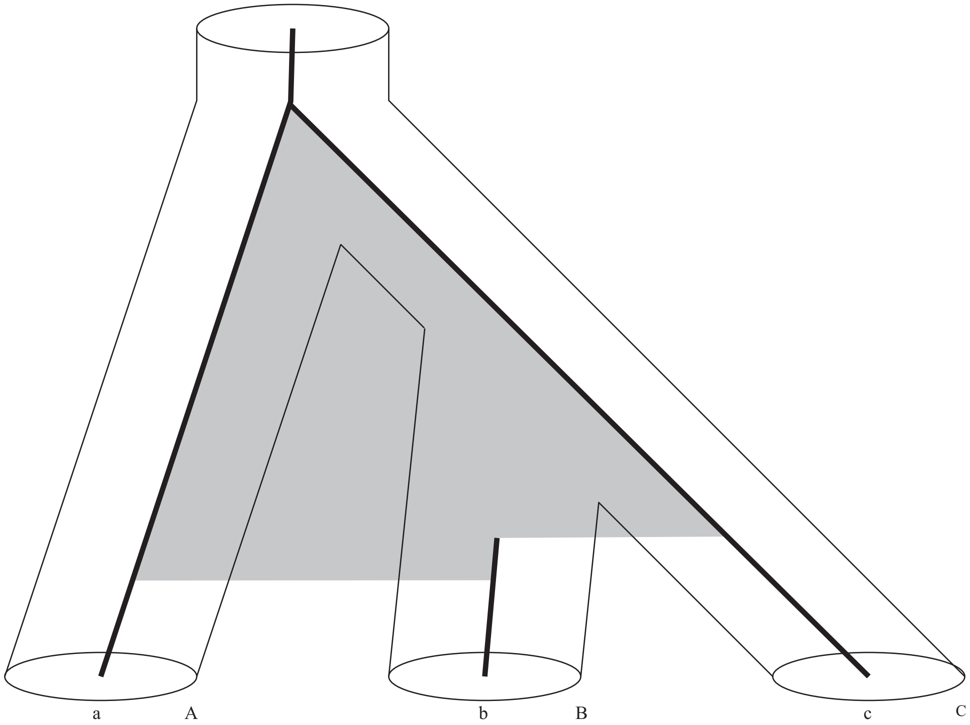


Figure 3: A parasite fails to speciate and then follows after host's speciation. These events are described with notation in [Pages, 2003].

Example

Example: $k = 1$ and $n = 4$

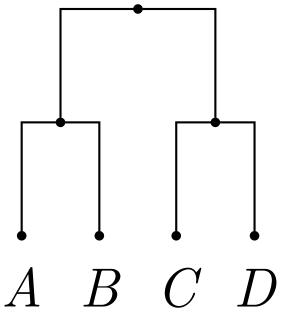
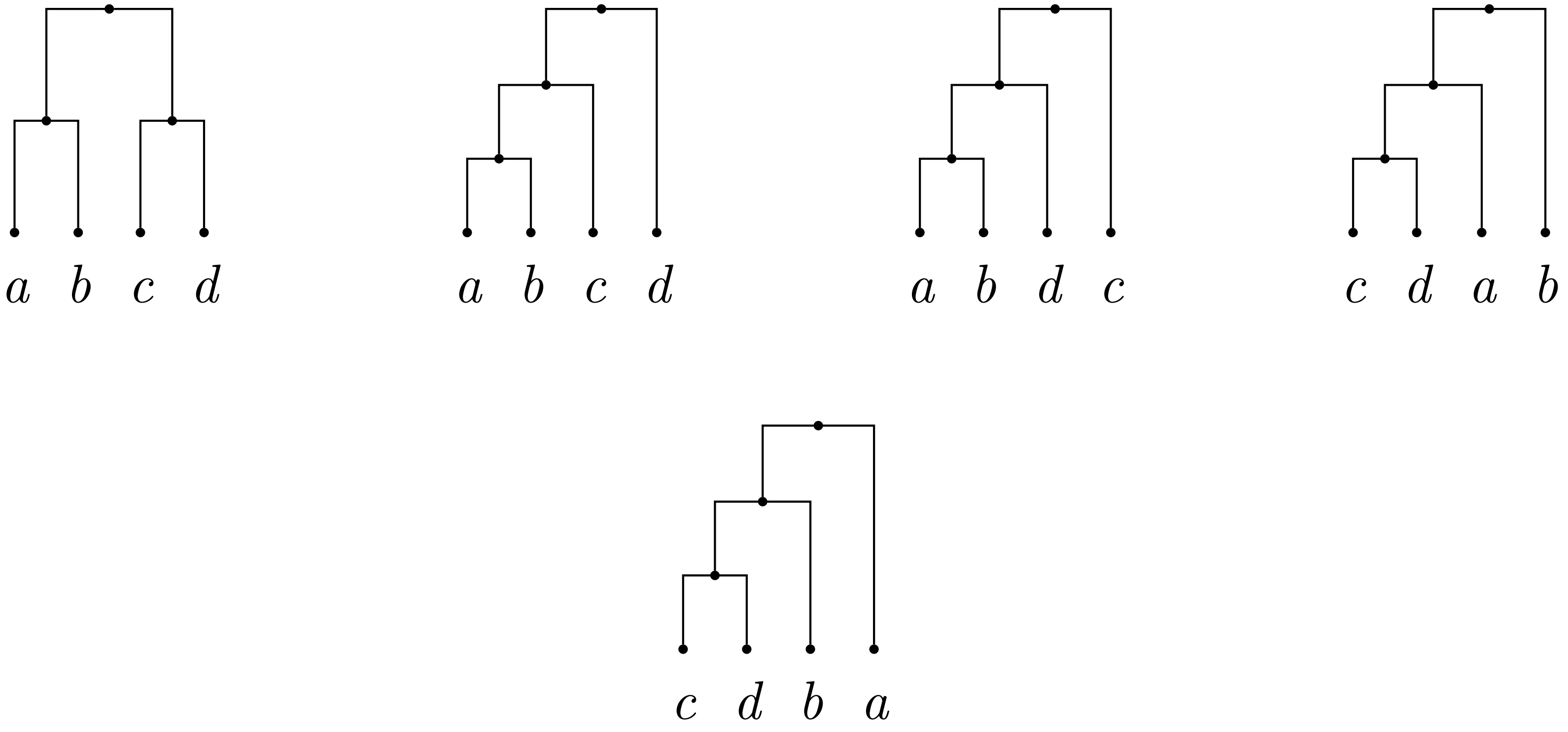


Figure 4: Host tree.

Example...

There are 5 possible parasite tree topologies.



Open Problems

Cophylogenetic invariants

Phylogenetic invariants are a well-stuided subject in [ASCB], and can be generalized to cophylogeny.

Fix a group-based model for gene sequence evolution. Suppose we know a species tree (or a host tree) T_H and we assume that gene trees have to be similar to the species tree (e.g., within a prescribed NNI distance or k-interval).

Consider the ideal of invariants I_{T_P} of phylogenetic invariants for each compatible gene-tree topology $T_P \in S_{T_H}$.

The intersection of these ideals (over all gene trees compatible with the species tree) gives invariants which describe gene-species tree compatibility.

Question: Can we describe/understand some generators of the intersection ideal, in terms of the original species tree – without resorting to a brute force computation of the intersection ideal?

Space of trees

Question: Is there any interesting space of cophylogenetic trees which can be described geometrically?

Example: The "topology diagonal"

Let $S = \{(D_H, D_P) : D_H$ is a tree metric for T_H and D_P is a tree metric for T_P such that T_H, T_P have the same tree topology$\}$.

S can be described by an extended four-point condition: the four-point condition has to hold for D_H, D_P, and $D_H + D_P$.

Constrained phylogenetic reconstruction

If host/parasite trees are reconstructed independently, then the disagreement between the reconstructed trees is exaggerated, because disagreement was not penalized during the reconstruction.

Question 1: Want to formulate minimum evolution or ML reconstruction methods which include penalties for co-evolution events that change tree topology?

Question 2: Want to develop distance-based methods for fast joint reconstruction, and we would like to understand them geometrically. One could also formulate projecting a pair of dissimilarity maps (D_H, D_P) onto a constrained space of cophylogenetic trees.

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.

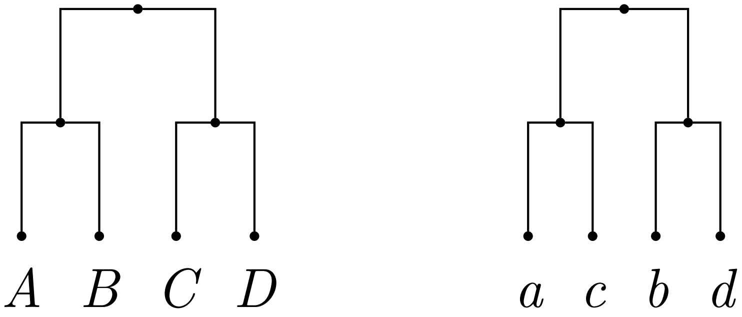
If we find multiple clusters, then they might represent recombinations.

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 kernel 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, 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.

Statistical tests

We want to test if host/parasite trees look “significantly different”, i.e. our hypotheses are

H_0: Host and parasite sequences are co-evolved according to the same tree topology in terms of a given kernel.

H_1: The true host/parasite tree topologies are not the same, due to host-switching etc.

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.

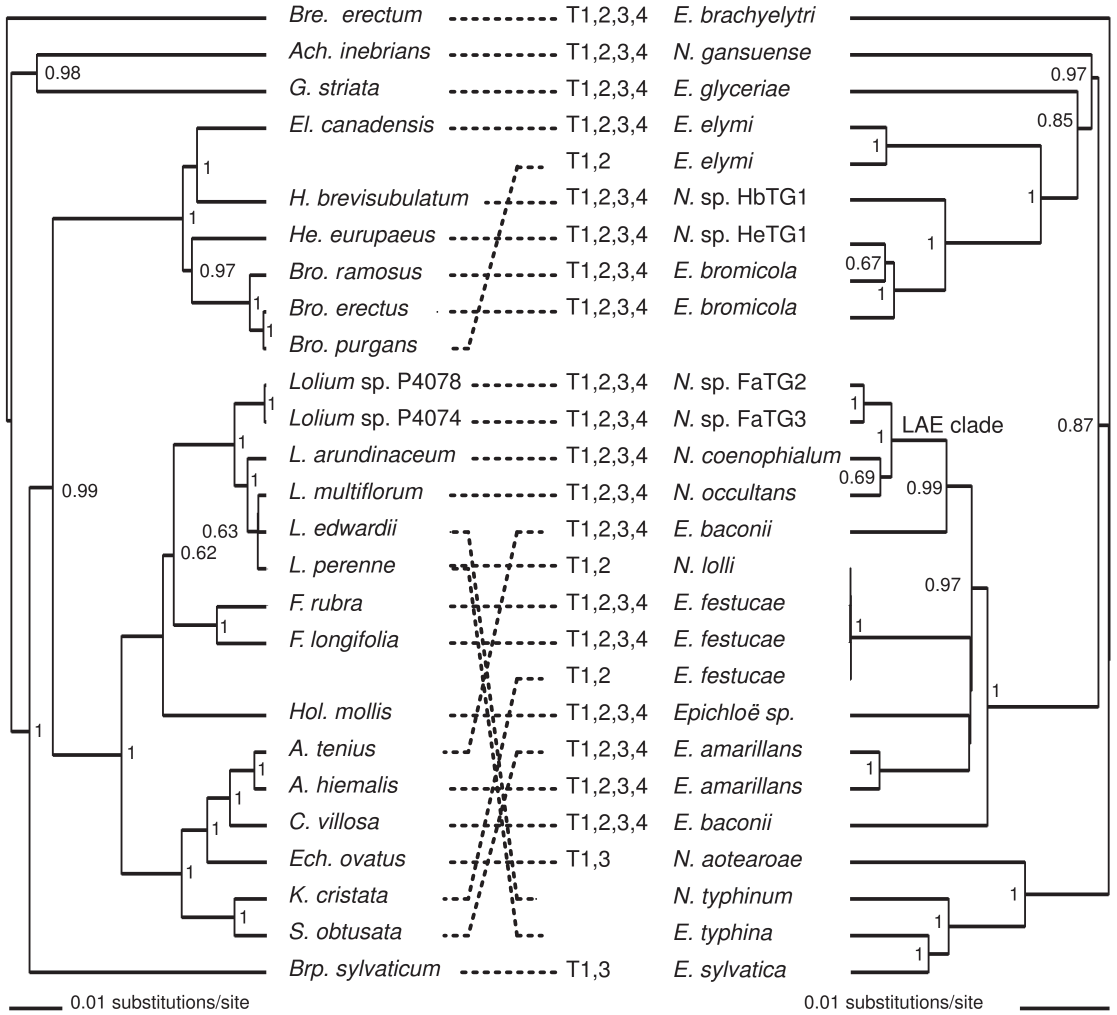


Figure 5: Ultrametric ML time trees for plant and endophyte data sets in [Schardl et al, 2008] constructed via BEAST. Hosts and their endophytes are indicated by dashed lines. Numeric values on nodes represent their posterior probabilities estimated by BEAST.

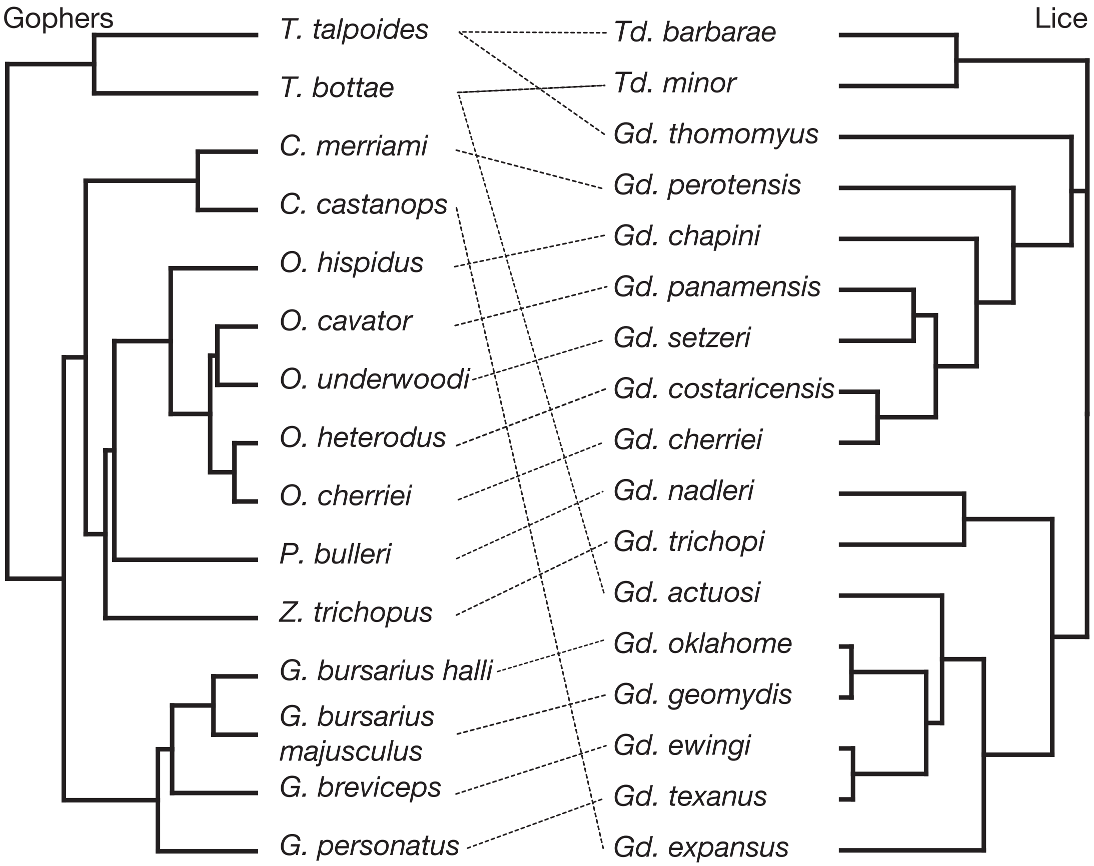


Figure 6: Ultrametric ML time trees for gopher and louse data sets [Hafner, 1990] constructed via BEAST. [Page and Hafner, 1994] and [Huelsenbeck et. al., 2000] studied these data sets.

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"?

Thank you....

http://arxiv.org/abs/0809.1908

Supported by NIH