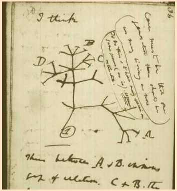
Tree Space: A historical perspective.

Susan Holmes
Banff, February 13, 2017 (Darwin+1)

Bio-X and Statistics, Stanford University

NIH-R01GM086884

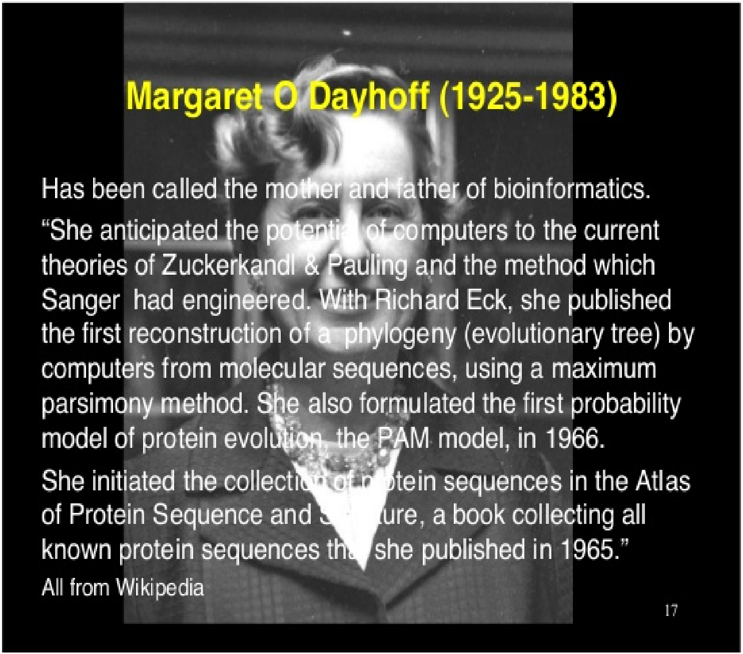


Phylogenetics: ‘some’ pioneers: 1970s-1980’s.

Geneticists

Statisticians and Probabilists

Mathematicians:



Margaret O Dayhoff (1925-1983)

Has been called the mother and father of bioinformatics.

“She anticipated the potential of computers to the current theories of Zuckerkandl & Pauling and the method which Sanger had engineered. With Richard Eck, she published the first reconstruction of a phylogeny (evolutionary tree) by computers from molecular sequences, using a maximum parsimony method. She also formulated the first probability model of protein evolution, the PAM model, in 1966.

She initiated the collection of protein sequences in the Atlas of Protein Sequence and Structure, a book collecting all known protein sequences that she published in 1965.”

All from Wikipedia

Statistical Inference: 1980s-90s

- Efron’s bootstrap, 1979.

Felsenstein suggests the Bootstrap for Phylogenies in 1985

Bootstrap support for Phylogenies.

Taking as observations the *columns* of the matrix X of aligned sequences, the rows representing the species.

The sampling distribution of the estimated tree is estimated by resampling with replacement among the characters or columns of the data.

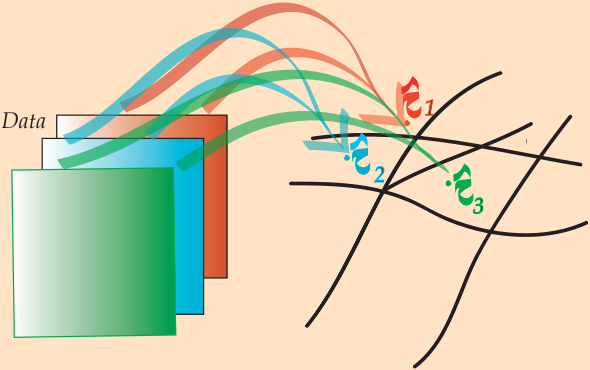
This provides a large set of plausible alternative data sets, each be used in the same way as the original data to give a separate tree. (see **?** for a review).

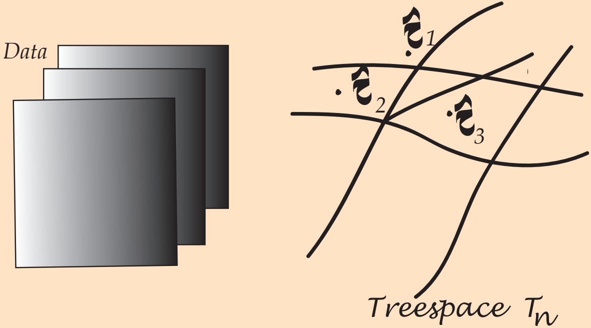
Simple confidence values

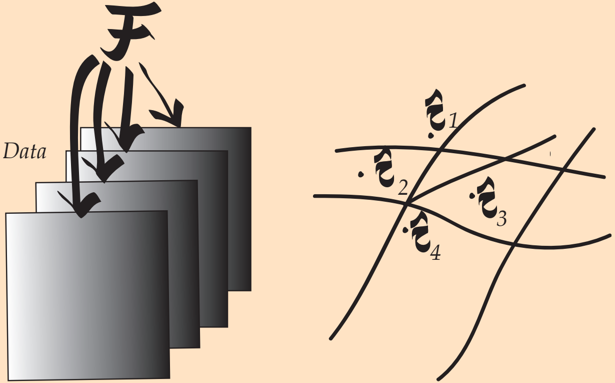
- Univariate.
- Multiple Testing.
- Composite Statements.

```
                                   +----Macaca mul
                              +-100.0
                         +-99.5    +----Macaca fus
                         !    !
               +------100.0   +---------Macaca fas
               !         !
               !         +--------------Macaca syl
          +-79.4
          !    !    +-------------------Hylobates
          !    !    !
          !    !    !              +----Homosapien
          !    +-99.0         +-50.2
     +-100.0        !    +-100.0   +----Pan
     !    !         !    !    !
     !    !         +-89.0    +---------Gorilla
     !    !              !
+-100.0   !              +--------------Pongo
!    !    !
!    !    +-----------------------------Saimiri sc
!    +----------------------------------Tarsius sy
+---------------------------------------Lemur catt
```

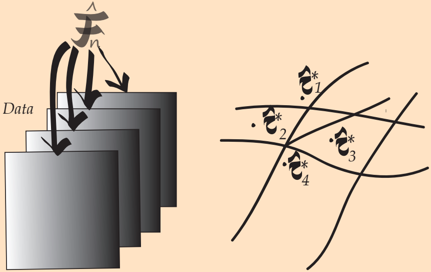
Sampling Distribution for Trees




Data
$\mathfrak{T}_1$
$\mathfrak{T}_2$
$\mathfrak{T}_3$
Treespace T_n



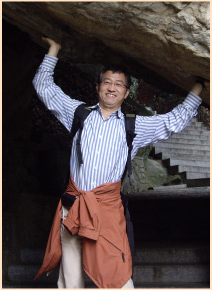
True Sampling Distribution



Bootstrap Sampling Distribution
(non parametric)

Statistical Inference: 1990s

- Bayesian methods using MCMC.
  Ziheng Yang, Bret Larget, Michael Newton, Hani Doss, Bruce Rannala, John Hulsenbeck.



Early 1990's

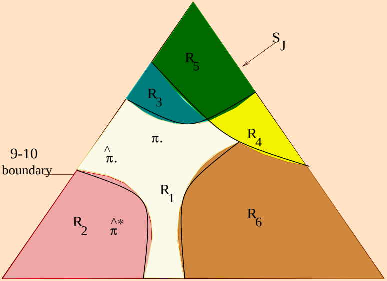
Stanford

Michael Newton came and gave a talk on **Large deviations for the bootstrap for trees**.

Efron, Halloran, H. , (1996)

Bootstrap confidence levels for trees

Depend on local and global properties of a neighborhood.

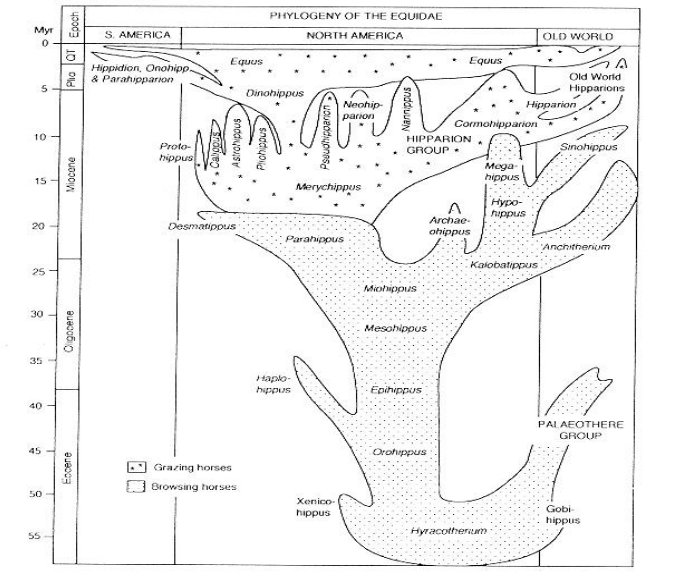


From Efron, Halloran, H., (1996)

What is the curvature of the boundary?

How many neighbors does a region have?

Confidence Statements for trees



Bootstrap for Multidimensional Scaling and PCA (H.,1985)

Schoenberg's (1935) remarked that a symmetric matrix of positive entries with zeros on the diagonal is a Euclidean distance matrix between n points if and only if the matrix

$$-\frac{1}{2}H\Delta_2 H \text{ is semi-definite positive}$$

where $H = (I - \frac{1}{n}\mathbf{1}\mathbf{1}')$, and $\mathbf{1}' = (1, 1, 1 \ldots, 1)$

Approximating Non Euclidean Distances by Euclidean ones

Forward:Decomposition of Distances Suppose we did have an Euclidean space, variables measured in $\mathbb{R}^p$ that are not centered: Y, apply the centering matrix

$$X = HY, \qquad \text{with } H = (I - \frac{1}{n}\mathbf{1}\mathbf{1}'), \text{and } \mathbf{1}' = (1, 1, 1 \ldots, 1)$$

Call $B = XX'$, if $D^{(2)}$ is the matrix of squared distances between rows of X in the euclidean coordinates,

$$d_{i,j} = \sqrt{(x_i^1 - x_j^1)^2 + \cdots + (x_i^p - x_j^p)^2}. \text{ and } -\frac{1}{2}HD^{(2)}H = B$$

Backward from D to X We can go backwards from a matrix D to X by taking the eigendecomposition of B in much the same way that PCA provides the best rank r approximation for data by taking the singular value decomposition of X, or the eigendecomposition of XX'.

$$X^{(r)} = US^{(r)}V' \text{ with } S^{(r)} = \begin{pmatrix} s_1 & 0 & 0 & 0 & \dots \\ 0 & s_2 & 0 & 0 & \dots \\ 0 & 0 & \dots & \dots & \dots \\ 0 & 0 & \dots & s_r & \dots \\ \dots & \dots & \dots & 0 & 0 \end{pmatrix}$$

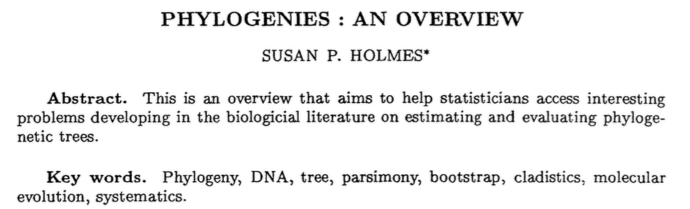
This provides the best approximate representation in an Euclidean space of dimension r. The algorithm provides points in a Euclidean space that have approximately the same distances as those provided by D^2.

MDS Algorithm

In summary, given an $n \times n$ matrix of interpoint distances, one can solve for points achieving these distances by:

1. Double centering the interpoint distance squared matrix: $S = -\frac{1}{2} H D_2 H$.
2. Diagonalizing S: $S = U \Lambda U^T$.
3. Extracting $\tilde{X}$: $\tilde{X} = U \Lambda^{1/2}$.

IMA Workshop: 1996



PHYLOGENIES : AN OVERVIEW

SUSAN P. HOLMES*

Abstract. This is an overview that aims to help statisticians access interesting problems developing in the biologicial literature on estimating and evaluating phylogenetic trees.

Key words. Phylogeny, DNA, tree, parsimony, bootstrap, cladistics, molecular evolution, systematics.

The phylogenetic tree is a statistical parameter, estimated in different ways from DNA/AA data:

- Parametric: ML estimation, PAML, Phyml, FastML,RaxML,...
- Distance based methods: Neighbor Joining, UPGMA,..
- Parsimony: Steiner tree problem: nonparametric.
- Bayesian estimation, Mr Bayes by MCMC, from posterior sampling distribution.

MIT Combinatorialist

Felsenstein, 1978 had published the number of phylogenetic trees

$$(2n-3)!! = (2n-3) \times (2n-5) \times \ldots 5 \times 3$$

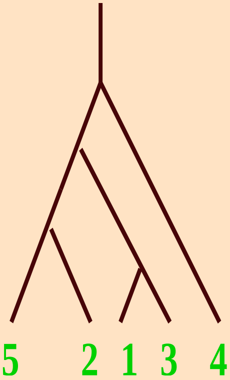
This formula for the number of trees was first proved using generating functions by Schroder (1873)**?**.

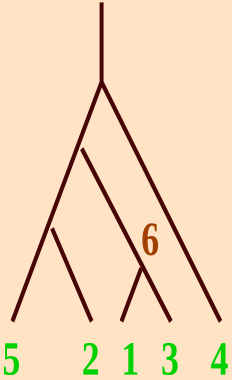
Coding Trees as Perfect Matchings

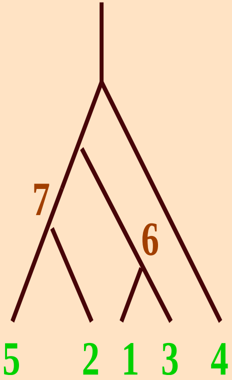
A perfect matching on $2n$ points is a partition of $1, 2, \ldots, 2n$ into n two-element subsets. It is well known that there are $(2n)!/2^n n!$ distinct perfect matchings. When $n = 2$, the three perfect matchings are

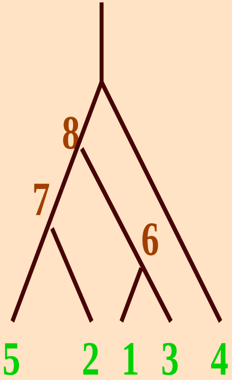
$$\{1, 2\}\{3, 4\}; \{1, 3\}\{2, 4\}; \{1, 4\}\{2, 3\}$$

From Trees to Matchings




6
5
2
1
3
4

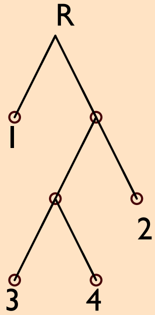

7
6
5
2
1
3
4



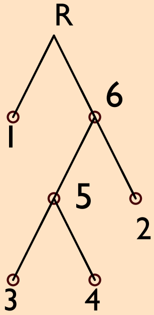
Put down the sibling pairs:

$$(1,3)(2,5)(6,7)(8,4)$$

We briefly describe the correspondence between matchings and trees. Begin with a tree with ℓ labeled leaves. Label the internal vertices sequentially with $\ell+1, \ell+2, \ldots, 2(\ell-1)$ choosing at each stage the ancestor which has both children labeled and who has the descendent lowest possible available label (youngest child). Thus the tree



is labeled



.

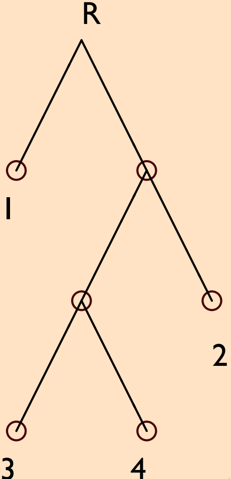
When all nodes are labeled, create a matching on $2n = 2(\ell - 1)$ vertices by grouping siblings. In the example above, this yields

$$\{3,4\}\{2,5\}\{1,6\}.$$

From matchings to trees

To go backward, given a perfect matching of $2n$ points, note that at least one matched pair has both entries from $\{1, 2, 3 \ldots, n+1\}$. All such labels are leaves; if there are several leaf-labeled pairs, choose the pair with the smallest label. Give the next available label ($n+2 = \ell+1$) to their parent node. There are then a new set of available labeled pairs. Choose again the pair with the smallest label to take the next available label for its parent, and so on.

For example, $\{3,4\}\{2,5\}\{1,6\}$ has $2n = 6$ and $\{3,4\}$ has both entries from $\{1,2,3,4\}$. The parent of these is labeled 5 and thus matched with 2 and then the parent of $\{2,5\}$ is matched with 1, yielding



Matchings and Decompositions

Diaconis and Holmes (1998) A matching of 2(n-1) objects is a pairing off, without care for order within pairs or between pairs.

The Same matchings:

$$(1,4)(2,5)(3,6)$$
$$(6,3)(4,1)(2,5)$$
$$(5,2)(3,6)(1,4)$$

Call $\mathcal{B}_{n-1}$ the subgroup of $\mathcal{S}_{2n-2}$ that fixes the pairs

$$\{1,2\}\{3,4\}\dots\{2n-3,2n-2\}$$

then

$$\mathcal{M}_{n-1} = \mathcal{S}_{2n}/\mathcal{B}_{n-1}$$

and

$$|\mathcal{M}_{n-1}| = \frac{(2n-2)!}{2^{n-1}(n-1)!} = (2n-3)!! = (2n-3)\times(2n-5)\times\cdots\times 3\times 1$$

$(\mathcal{S}_{2n-2}, \mathcal{B}_{n-1})$ form a Gelfand pair Diaconis and Shahshahani (1987)

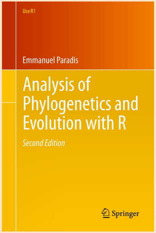
$$L(\mathcal{M}_{n-1}) = V_1 \oplus V_2 \oplus \ldots \oplus V_\lambda$$

A multiplicity free representation.

$$L(\mathcal{M}_{n-1}) = \bigoplus_{\lambda \vdash n} \mathcal{S}^{2\lambda}$$

where the direct sum is over all partitions λ of m, $2\lambda = (2\lambda_1, 2\lambda_2, \ldots, 2\lambda_k)$ and $\mathcal{S}^{2\lambda}$ is associated irreducible representation of the symmetric group S_{2m}.
Just to take the first few: for $\lambda = n - 1$ S^λ are the constants, and this gives the sample size. for $\lambda = (n-2, 1), S^\lambda$ are the number of times each pair appears. for $\lambda = (n-3, 2), S^\lambda$ are the number of times partition of 4 appears in the tree. for $\lambda = (n-3, 1, 1), S^\lambda$ are the number of times 2 pairs appear simultaneously.

Matchings are useful



Use R!

Emmanuel Paradis

Analysis of Phylogenetics and Evolution with R

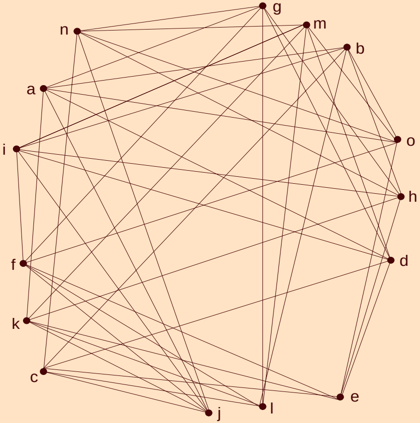
Second Edition

Springer

- For going through all trees systematically. (Gray code for Trees)
- Doing vigorous random walks on tree space.
- Doing Fourier Analysis on Tree Data.

But the matching distance is not satisfactory to the biologists.

The Matching Polytope



The Gelfand pair decomposition is similar to what was done by Diaconis for permutation data (IMS Book on Probability and Statistics on Groups, 1988).

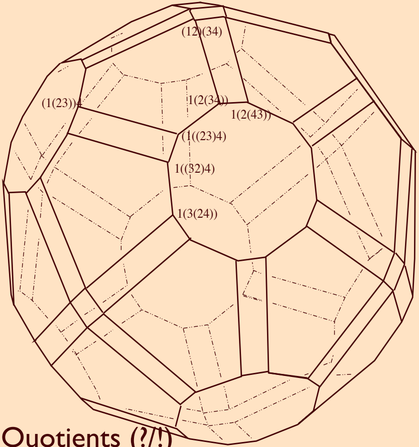
Mallow's Model

$$P(\tau) \quad = \quad Ke^{-\lambda d(\tau,\tau_0)}, \qquad \text{K is a normalizing constant}$$

- Exponential family, it needs:
  - A central tree τ_0
  - A distance between trees $d(\tau, \tau_0)$
- It is possible to extend this to make a Bayesian model with a symmetric apriori distribution for τ_0.

Distances and centroids are essential

Cornell, 1997: The permuto-associahedron



Quotients (?/!)

A book on polytopes.(Ziegler)

But the trees are extreme points

Talk with Lou Billera:

Frequentist Confidence Regions

$$P(\tau \in \mathcal{R}_\alpha) = 1 - \alpha$$

We will use the nonparametric approach of Tukey who proposed peeling convex hulls to construct successive ‘deeper’ confidence regions. But we need a geometrical space to build these regions in.

What does a neighborhood look like?

Need modern topology.

Aims

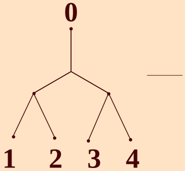
- Fill Tree Space and make meaningful boundaries.
- Define distances between trees.
- Define neighborhoods, meaningful measures.
- Principal directions of variations in tree space, summarizing : structure + noise.
- Confidence statements, convex hulls.

Distances between Trees

Karen Vogtmann

- Robinson and Foulds, (bipartitions).
- Nearest Neighbor Interchange (NNI).

  Rotation Moves

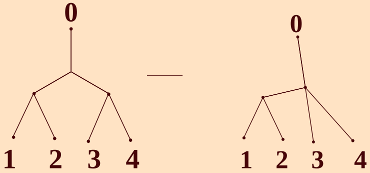


Distances between Trees

Karen Vogtmann

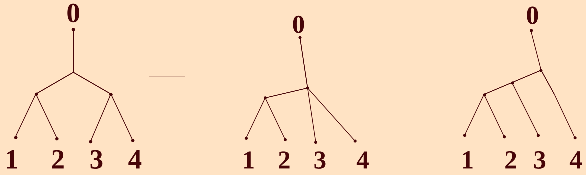
- Robinson and Foulds, (bipartitions).
- Nearest Neighbor Interchange (NNI).

  Rotation Moves



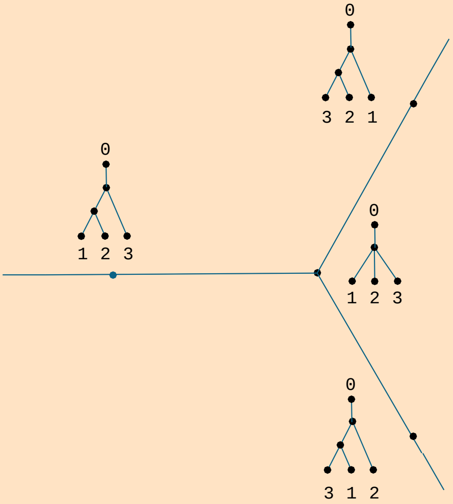
Distances between Trees Karen Vogtmann

- Robinson and Foulds, (bipartitions).
- Nearest Neighbor Interchange (NNI).
  Rotation Moves

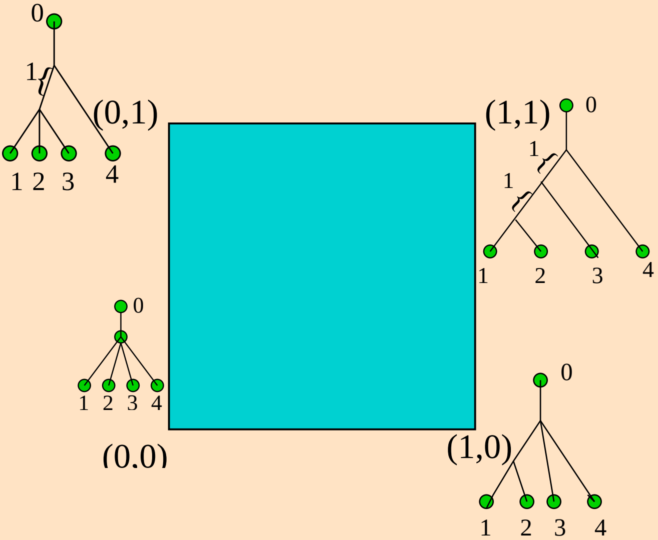


- Fill-in of NNI moves: Billera, Holmes, Vogtmann (BHV).
  The boundaries between regions represent an area of uncertainty about the exact branching order. In biological terminology this is called an 'unresolved' tree.

Boundary for trees with 3 leaves

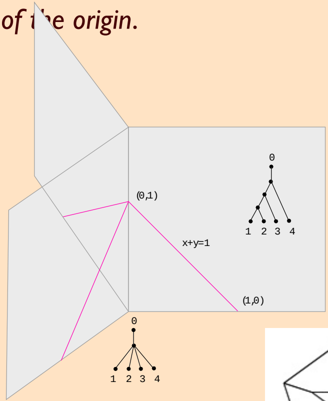


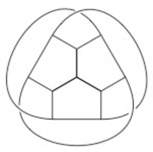
The quadrant for one tree

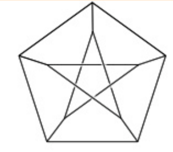


Link of the origin

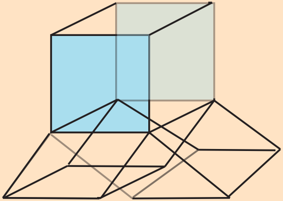
All 15 quadrants for $n = 4$ share the same origin. If we take the diagonal line segment $x + y = 1$ in each quadrant, we obtain a graph with an edge for each quadrant and a trivalent vertex for each boundary ray; this graph is called the *link of the origin*.







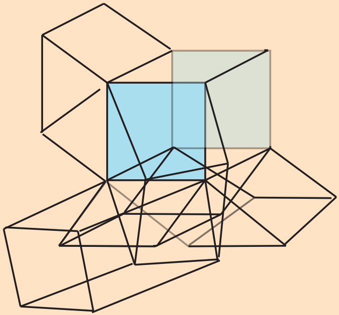
Cube complex of Euclidean Orthants



A path between two trees consists of line segments through a sequence of orthants. This sequence of orthants is the *path*.

A path is a *geodesic* when it has the smallest length of all paths between two points.

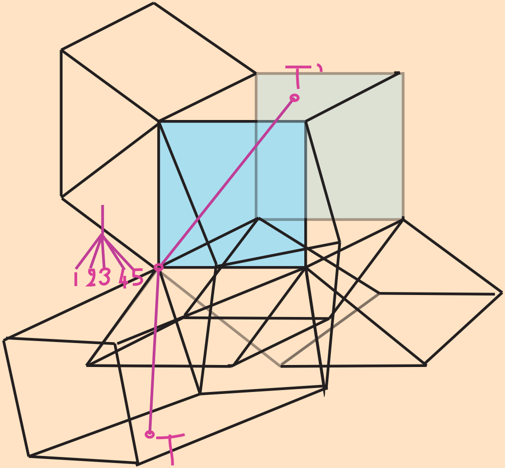
Cube complex of Euclidean Orthants



A path between two trees consists of line segments through a sequence of orthants. This sequence of orthants is the *path*.

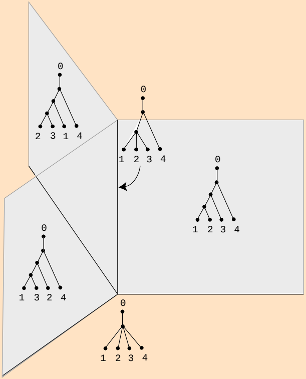
A path is a *geodesic* when it has the smallest length of all paths between two points.

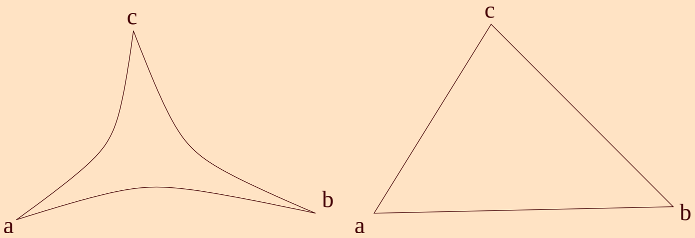
A Cone Path



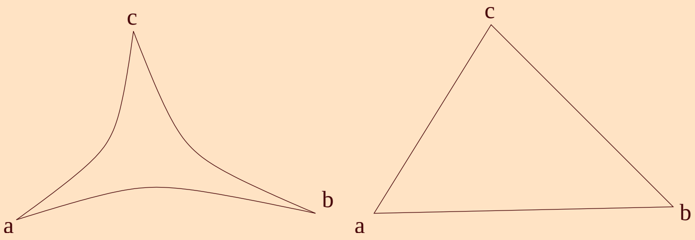
A path between two trees T and T′ always exists. Since all orthants connect at the origin, any two trees T and T′ can be connected by a two-segment path, this is called the cone-path.

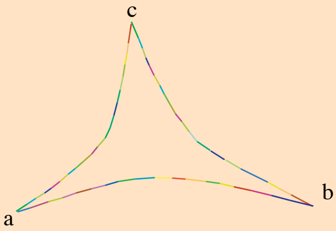
Three orthants sharing a common boundary for $n = 4$ leaves.

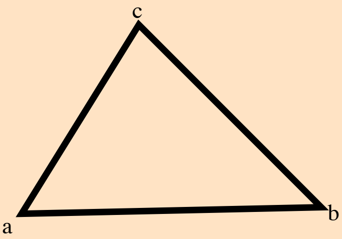




Theorem(Billera, Holmes, Vogtmann (BHV)): Tree space with BHV metric is a CAT(0) space, that is, it has non-positive curvature. This implies there are geodesic between any two trees (Gromov). It is not an Euclidean space.


c
a
b
c
a
b


c
b
a


c
a
b

This has an effect on the existence of geodesics.

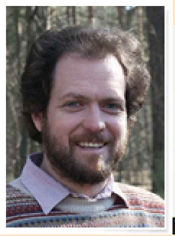
The speed at which MCMC methods work.

The size of the “variance”.

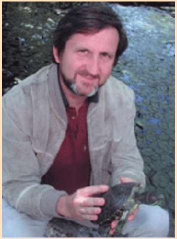
The computation of the mean of a set of trees.

The number of neighbors of a tree.

2001: NZ Phylogenetics



KITP: May 2001: David Hillis



Santa Barbara: I taught him all about Multidimensional Scaling and he put me in contact with some biologists with difficult multiple tree source problems.

AIM conference 2002, Palo Alto

Francis Su has extensive notes:
Felsenstein, Billera, Vogtman, H., Wachs, Diaconis, Shaharian, Staple, Vert, K. St John, Nina Amenta, David Epstein.
.. Further progress on the right way to encode trees through a binary encoding of edges
Use edge compatibility to go through treespace.

We can embed trees in Euclidean space (approximately) using MDS

Geodesic metric space:

If we have a distance defined between any two points of a space, we call it a metric space.

(The distance doesn't have to be defined through ordinary coordinates)

A geodesic metric space is a metric space where geodesics are defined to be the shortest path between points in the space.

We can ask whether points are closer to a tree or to being embeddable in Euclidean space by using Gromov's δ.

Implementation:

`distory` is an `R` package written with John Chakerian**?** which both implements the geodesic BHV distance between trees using Owen and Provan (2009)'s algorithm and the computation of delta for any finite set of points.

We know that given a distance matrix we can give a treelike representation of the points with these distances by building a tree if the distances obey Buneman's four point condition (Buneman, 1974).

Buneman's four point condition

For any four points (u, v, w, x) :
The three sums:$d(u, v) + d(w, x)$, $d(u, w) + d(v, x)$, $d(u, x) + d(v, w)$ are equal, not less than the third.

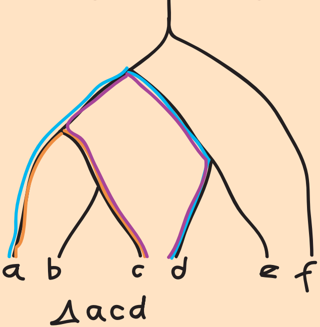
We can see Gromov's definition the hyperbolicity constant δ as a relaxation of the above four-point condition:

Gromov's hyperbolicity contant

For any four points u,v,w,x, the two larger of the three sums $d(u,v)+d(w,x), d(u,w)+d(v,x), d(u,x)+d(v,w)$ differ by at most 2δ.

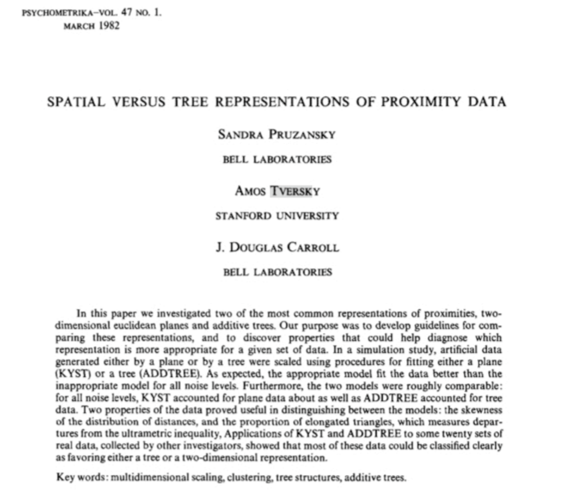
δ-hyperbolic space is a geodesic metric space in which every geodesic triangle is δ-thin.
δ-thin: pick three points and draw geodesic lines between them to make a geodesic triangle. Then any point on any of the edges of the triangle is within a distance of δ from one of the other two sides.

For example, trees are 0-hyperbolic: a geodesic triangle in a tree is just a subtree, so any point on a geodesic triangle is actually on two edges.



Normal Euclidean space is ∞-hyperbolic; i.e. not hyperbolic. Generally, the higher δ has to be, the less curved the space is.

Is it better to represent the distances by a tree or a Euclidean projection?



PSYCHOMETRIKA–VOL. 47 NO. 1.
MARCH 1982

SPATIAL VERSUS TREE REPRESENTATIONS OF PROXIMITY DATA

SANDRA PRUZANSKY

BELL LABORATORIES

AMOS TVERSKY

STANFORD UNIVERSITY

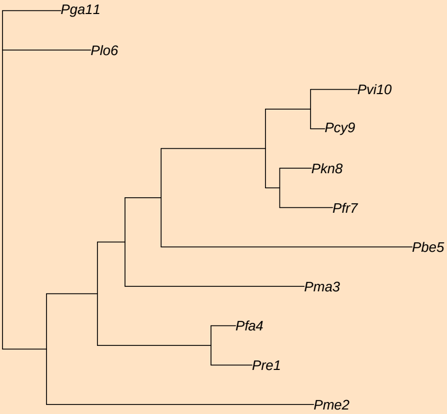
J. DOUGLAS CARROLL

BELL LABORATORIES

In this paper we investigated two of the most common representations of proximities, two-dimensional euclidean planes and additive trees. Our purpose was to develop guidelines for comparing these representations, and to discover properties that could help diagnose which representation is more appropriate for a given set of data. In a simulation study, artificial data generated either by a plane or by a tree were scaled using procedures for fitting either a plane (KYST) or a tree (ADDTREE). As expected, the appropriate model fit the data better than the inappropriate model for all noise levels. Furthermore, the two models were roughly comparable: for all noise levels, KYST accounted for plane data about as well as ADDTREE accounted for tree data. Two properties of the data proved useful in distinguishing between the models: the skewness of the distribution of distances, and the proportion of elongated triangles, which measures departures from the ultrametric inequality, Applications of KYST and ADDTREE to some twenty sets of real data, collected by other investigators, showed that most of these data could be classified clearly as favoring either a tree or a two-dimensional representation.

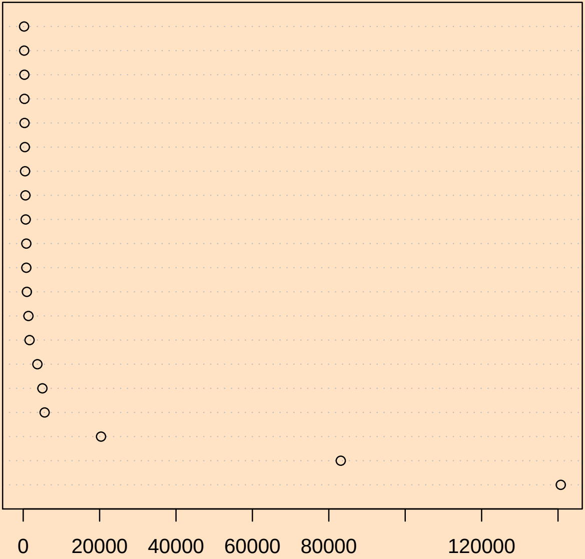
Key words: multidimensional scaling, clustering, tree structures, additive trees.

Malaria Data as seen using ape

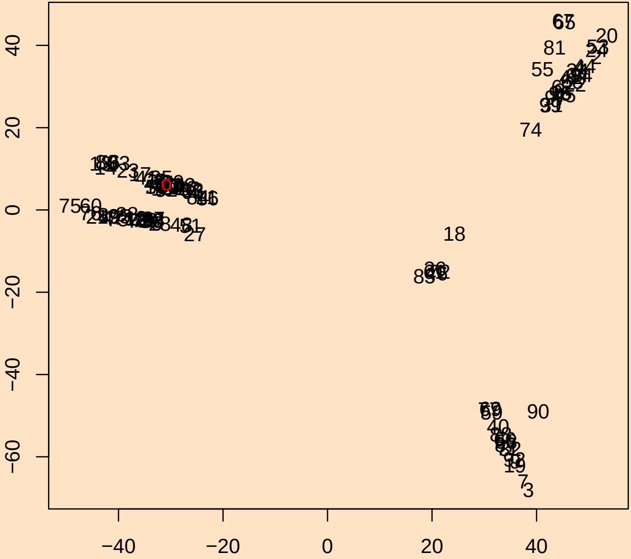


Bootstrap of Malaria Data

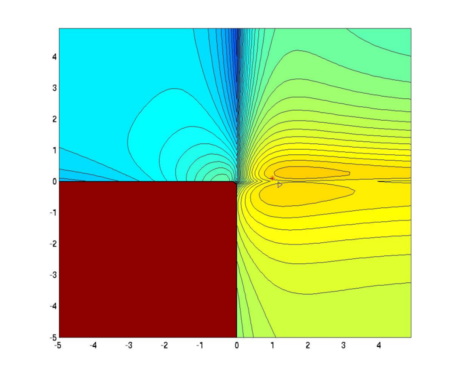
Eigenvalues of MDS for bootstrapped trees



Bootstrapped trees

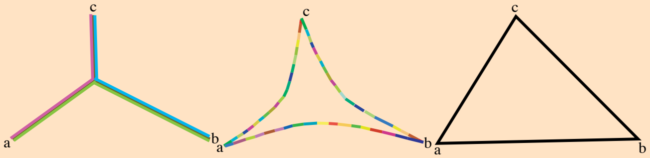


Maximum Likelihood Bootstrap



Tree of Trees

A tree is a complete CAT(0) space.



Since BHV,2001 **?** have shown that the space of trees is negatively curved (a **CAT(0)** space), the most natural representation of a collection of trees may be a tree.

Is this good for anything?

Statistical Uses for Distances

- Center of Cloud of Trees (equal weights): Find T_0 that minimizes either $\sum_{k=1}^{K} d^2(T_0, T_k)$ this is the (L^2) definition of the mean tree, or $\sum_{k=1}^{K} d(T_0, T_k)$ (L^1).
- Extend the above to cater for a measure on treespace.

$$P(T) = Kexp(-\lambda d(T, T_0))$$

- Variability of the tree-points:
  Pseudovariance=$\frac{1}{K-1} \sum_{k=1}^{K} d^2(T_0, T_k) = \hat{s}^2$.
- Studentizing :

$$\frac{d(\hat{T}^*, \hat{T}_{obs})}{\hat{s}}$$

- Leverage of a position, as in leverage of an observation in regression.
- PCA with regards to Instrumental Variables- DPCOA. Explain a set of distances between trees by other distances between the same data.

Thinking like a Statistician....

and a geometer..

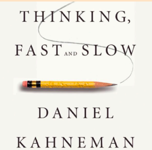
- How treelike are the data ? Model Selection and residual inspection: networks.
- Do we always need the tree, Distances between Data.
- Are all the characters supporting the tree? Leverage.
- Finding hidden gradients Ordination of trees.
- Stability under perturbation Evaluating the estimates.
- How variable are the trees? Variance and Moments.

Consequences

- Averaging works better than it should, (an argument against total evidence computation without decomposing??).
- We can build Bayesian priors based on distances.
- We can make a useful bootstrap statement.
- We can make convex hulls. $\longrightarrow$ Confidence regions.
- We know how many neighbors any tree has.
- We can make a useful bootstrap statement.

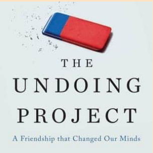
But distances are not everything....remember the baseline

Amos Tversky and Danny Kahnneman



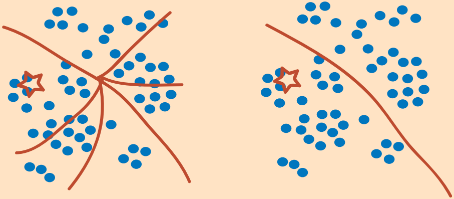
THINKING, FAST AND SLOW

DANIEL KAHNEMAN



.

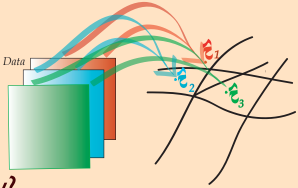
Heuristics and Biases, more particularly the representativeness heuristic.
Heuristics are described as ”judgmental shortcuts that generally get us where we need to go - and quickly - but at the cost of occasionally sending us off course.”
Heuristics are useful because they use effort-reduction and simplification in decision-making.
For representativenes of an event, similarity or a small distance is not enough, the baseline frequencies (ie probability) are essential to conclude.
We need careful realistic probability models for treespace, no real data has ever been uniform, no multivariate data is ever multivariate normal.



Beware the different number of neighbors matters if you think you are using a Monte Carlo method to estimate the distance to the boundary using the bootstrap.

Inferential Bootstrap: questions remain

$\mathcal{X}$ original data $\longrightarrow$ $\hat{\tau}$ estimate.



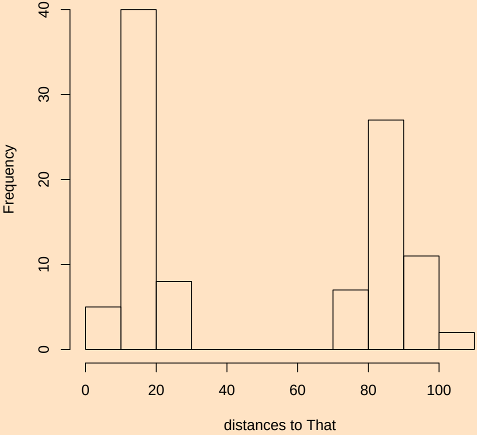
How?

Call $\mathcal{X}^*$ bootstrap samples consistent with the model used for estimating the tree:

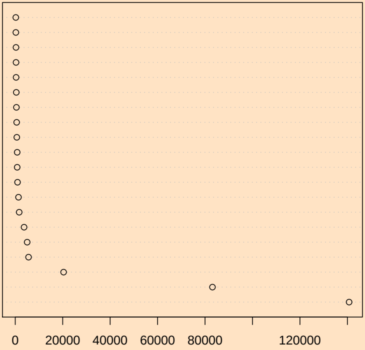
- Non parametric multinomial resampling for a parsimony tree.
- Seqgen parametric type resampling with the same parameters for a ML.
- Bayesian GAMMA prior on rates and generation (Yang 2000) for random sequences according to $\hat{\mathcal{T}}$

- Are the characters (columns) independent?
  We actually have less information than we think?
  What is the unit of information?
- Block Bootstrap to generate dependent data.
- Does the bootstrap work
  Conjecture:
  The bootstrap estimate of the sampling distribution of the distances $d(\hat{\tau}^*, \hat{\tau})$ is a good approximation to the true sampling distribution of $d(\hat{\tau}, \tau)$.

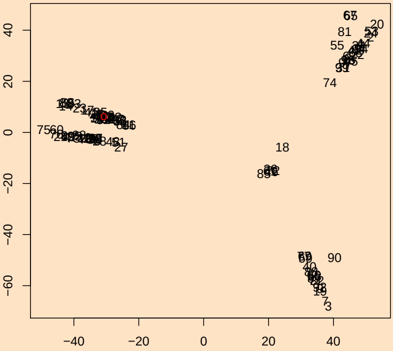
From Chakerian and H. (2011), using the algorithm by Owen and Provan (2010), implemented in C and wrapped into the `distory` package:



Eigenvalues of MDS for bootstrapped trees



Bootstrapped trees



Who Cares?

Bacterial Species in the Gut: Example of a Metagenome.

Samples from IBS and healthy rats give abundance of about 1,000 species of bacteria.

To be continued...

Benefitting from the tools and schools of Statisticians......

Thanks to the `R` community, in particular Robert Gentleman, Chessel, Thioulouse, Dray (`ade4` group in Lyon) and Emmanuel Paradis `ape`.
Collaborators: Louis Billera, Karen Vogtmann, Aaron Staple, Daniel Ford, John Chakerian, Persi Diaconis, Kris Sankaran, Elizabeth Purdom, Julia Fukuyama.
My website:
`http://webstat.stanford.edu/~susan/`
@SherlockpHolmes
or Google : `susan holmes stanford`

References

L. Billera, S. Holmes, and K. Vogtmann. The geometry of tree space. *Adv. Appl. Maths*, 771–801, 2001.

J. Chakerian and S. Holmes. Computational methods for evaluating phylogenetic trees, 2010. arXiv.

J. Chakerian and S. Holmes. distory:Distances between trees, 2010.

P. W. Diaconis and S. P. Holmes. Matchings and phylogenetic trees. *Proc. Natl. Acad. Sci. USA*, 95(25):14600–14602 (electronic), 1998.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979.

B. Efron, E. Halloran, and Susan P. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA*, 93:13429–34, 1996.

J. Felsenstein. *Inferring Phylogenies*. Sinauer, Boston, 2004.

M. Gromov. Hyperbolic groups. In *Essays in group theory*, pages 75–263. Springer, New York, 1987.

S. Holmes. Bootstrapping phylogenetic trees: theory and methods. *Statistical Science*, 18(2):241–255, 2003. Silver anniversary of the bootstrap.

S. Holmes. Statistical approach to tests involving phylogenies. In *Mathematics of Evolution and Phylogeny*. Oxford University Press, Oxford,UK, 2005.

R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

K. Mardia, J. Kent, and J. Bibby. *Multiariate Analysis*. Academic Press, NY., 1979.

E. Mossel. Phase transitions in phylogeny. *Trans. Amer. Math. Soc.*, 356(6):2379–2404 (electronic), 2004.

M. Owen and J.S. Provan. A fast algorithm for computing geodesic distances in tree space. *IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 2–13, 2010.

E. Paradis. Ape (analysis of phylogenetics and evolution) v1.8-2, 2006. `http://cran.r-project.org/doc/packages/ape.pdf`.

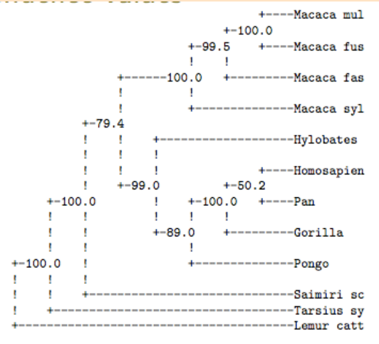
K Schliep. `phangorn`:Phylogenetic analysis in R, 2009.

I.J. Schoenberg. Remarks to Maurice Frechet's article "Sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'esp ace de Hilbert. *The Annals of Mathematics*, 36(3):724–732, July 1935.

E. Schröder. *Zeit. für. Math. Phys.*, 15:361–376, 1870.

F. H. Sheldon and A. H. Bledsoe. Avian molecular systematics. *Annu. Rev. Ecol. Syst.*, 24:243–278, 1993.

Comparing Different Trees

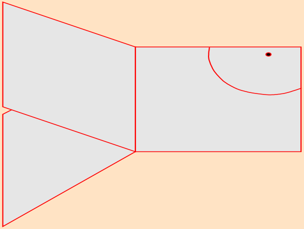


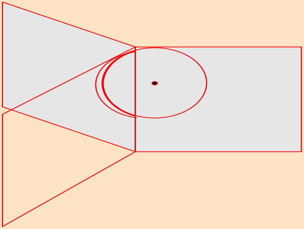
```
                                   +----Macaca mul
                              +-100.0
                         +-99.5    +----Macaca fus
                         !    !
               +------100.0   +---------Macaca fas
               !         !
               !         +--------------Macaca syl
          +-79.4
          !    !    +-------------------Hylobates
          !    !    !
          !    !    !              +----Homosapien
          !    +-99.0         +-50.2
     +-100.0        !    +-100.0   +----Pan
     !    !         !    !    !
     !    !         +-89.0    +---------Gorilla
     !    !              !
+-100.0   !              +--------------Pongo
!    !    !
!    !    +-----------------------------Saimiri sc
!    +----------------------------------Tarsius sy
+---------------------------------------Lemur catt
```

- Binomial Support Estimates (Consensus+support values).
- Split Differences, Visualization Programs .
- Distances.
- Recoding of Trees as binary columns.

How many neighbors for a given tree?(W.H.Li,1993)

We know the number of neighbors of each tree.



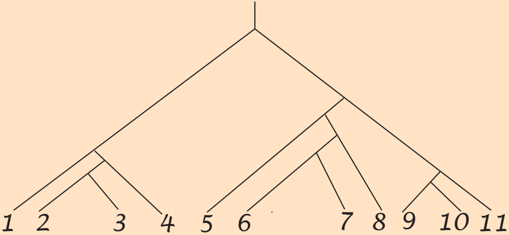


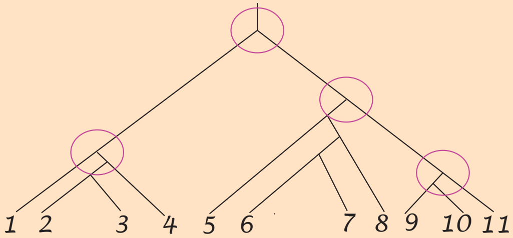
For a tree with only two inner edges, there is the only one way of having two edges small: to be close to the origin-star tree:
15 neighbors. This same notion of neighborhood containing 15 different branching orders applies to all trees on as many leaves as necessary but who have two contiguous “small edges” and all the other inner edges significantly bigger than 0.

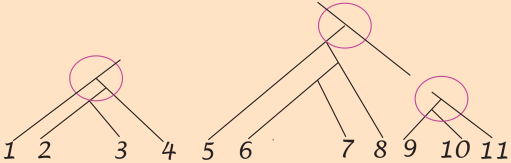
This picture of treespace frees us from having to use simulations to find out how many different trees are in a neighborhood of a given radius r around a given tree. All we have to do is check the sets of continguous edges in the tree smaller than r, say there is only one set of size k, then the neighborhood will contain

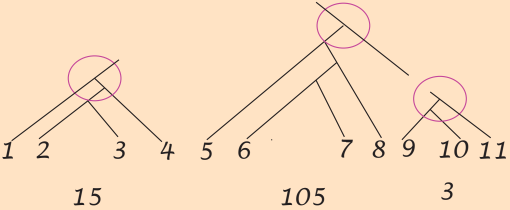
$$(2k-3)!! = (2k-3) \times (2k-5) \times \cdots 3\text{`different' trees.}$$

If there are m sets of sizes $(n_1, n_2, \ldots, n_m)$


1
2
3
4
5
6
7
8
9
10
11


1
2
3
4
5
6
7
8
9
10
11


1
2
3
4
5
6
7
8
9
10
11


1
2
3
4
5
6
7
8
9
10
11
15
105
3

In this case the number of trees within r will be $15 * 105 * 3 = 4725$, in general:

$$(2n_1 - 3)!! \times (2n_2 - 3)!! \times (2n_3 - 3)!! \cdots \times (2n_m - 3)!!$$

A tree near the star tree at the origin will have an exponential number of neighbors.
This explosion of the volume of a neighborhood at the origin provides for interesting math problems.

These differing number of neighbors for different trees show that the bootstrap values cannot be compared from one tree to another.

This was implicitly understood by Hendy and Penny in their NN Bootstrap procedure.

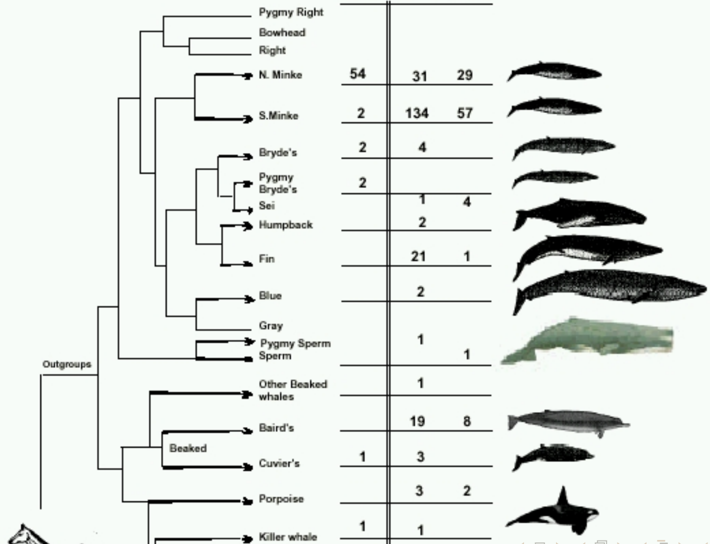
Are there other ways of using the bootstrap than just counting clade appearances?

Do we care about confidence statements for phylogenetic trees?

Cetacees: recognising what is being sold as Whale meat in Japan?

Steve Palumbi, Stanford, Scott Baker, Auckland.

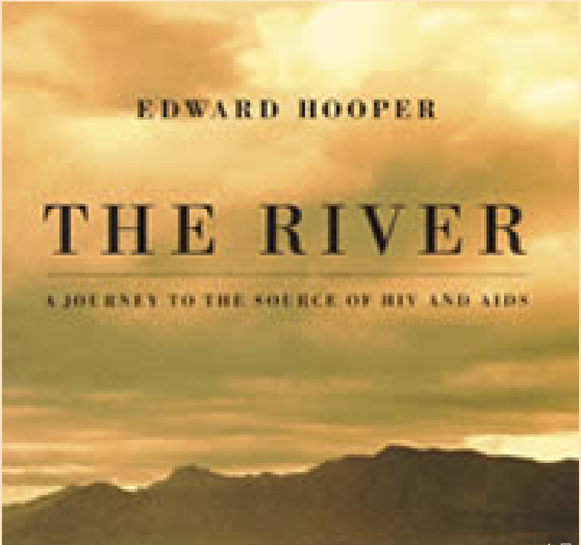
Phylogenetic Identification of Whale and Dolphin Products



	KOREA	JAPAN	
	1994-97	1993–98*	1998-99
Pygmy Right			
Bowhead			
Right			
N. Minke	54	31	29
S.Minke	2	134	57
Bryde's	2	4	
Pygmy Bryde's	2		
Sei		1	4
Humpback		2	
Fin		21	1
Blue		2	
Gray			
Pygmy Sperm		1	
Sperm			1
Other Beaked whales		1	
Baird's		19	8
Cuvier's	1	3	
Porpoise		3	2
Killer whale	1	1	

The River without a Paddle?

Human immunodeficiency virus: Phylogeny and the origin of HIV-1
The origin of human immunodeficiency virus type 1 (HIV-1) is controversial.



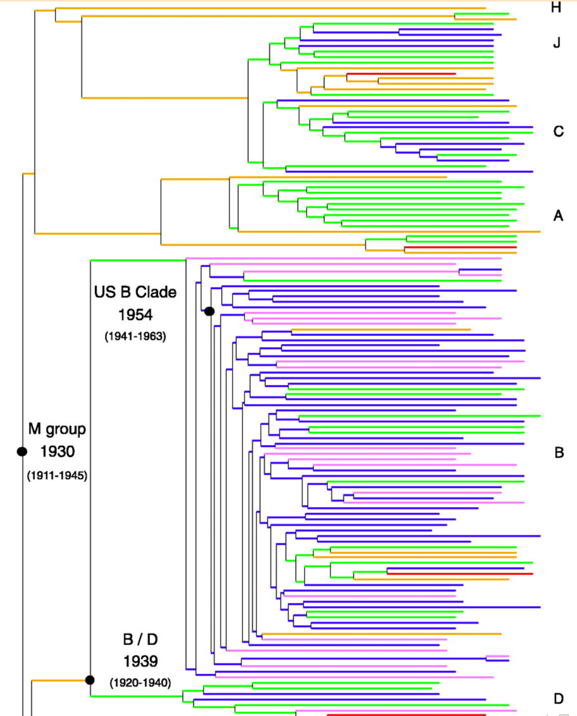
Conversely, phylogenetic analysis of HIV-1 sequences indicates that group M originated before the vaccination campaign, supporting a model of ’natural transfer’ from chimpanzees to humans. If this timescale is correct, then the OPV theory remains a viable hypothesis of HIV-1 origins only if the subtypes of group M differentiated in chimpanzees before their transmission to humans.

Confidence Intervals ?

Korber and colleagues extrapolated the timing of the origin of HIV-1 group M back to a single viral ancestor in 1931, give or take about 12 years for 95% confidence limits.

Because this calendar of events obviously pre-dated the OPV trials, in the revised version of his book, Hooper suggested that group M first began to diverge in chimpanzees, and that there were then several independent transfers of virus to humans via OPV.

In that case, several OPV batches should bear evidence of their production in chimpanzee tissue, yet no such evidence has been found.


H
J
C
A
US B Clade
1954
(1941-1963)
M group
1930
(1911-1945)
B
B / D
1939
(1920-1940)
D

Closure: Polio vaccines exonerated
Nature 410, 1035 - 1036 (2001)

The OPV batch that Hooper considered to be under most suspicion, however, was CHAT 10A-11.
An original vial of the batch was found at Britain’s National Institute for Biological Standards and Control, and the new tests show that it was prepared from rhesus-macaque cells.