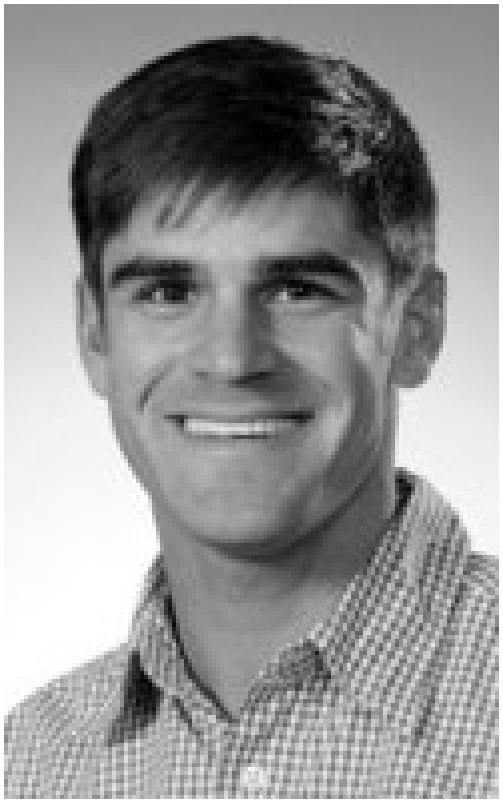
Metagenomics and metrics on spaces of probability measures

Steven N. Evans

Department of Mathematics & Department of Statistics
Group in Computational and Genomic Biology
Group in Computational Science and Engineering
University of California at Berkeley

March, 2011

Collaborator



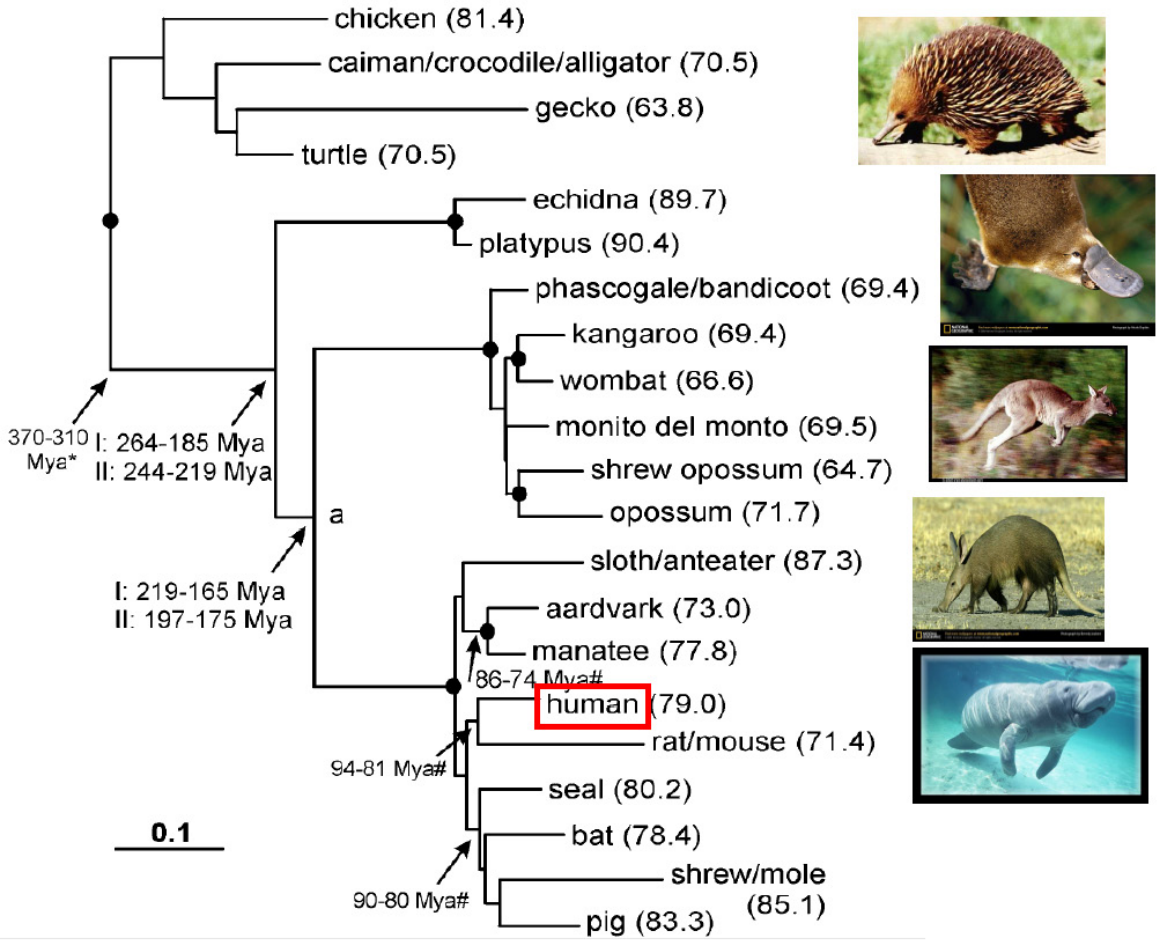
Erick Matsen (former post-doc)
Fred Hutchinson Cancer Research Center

Good advice from Erick

work with biologists who have interesting **data** *and who are nice people, too!*

What is phylogenetic inference?

In phylogeny we seek to reconstruct the evolutionary "family tree" of a number of (usually present day) taxa using data about those taxa.

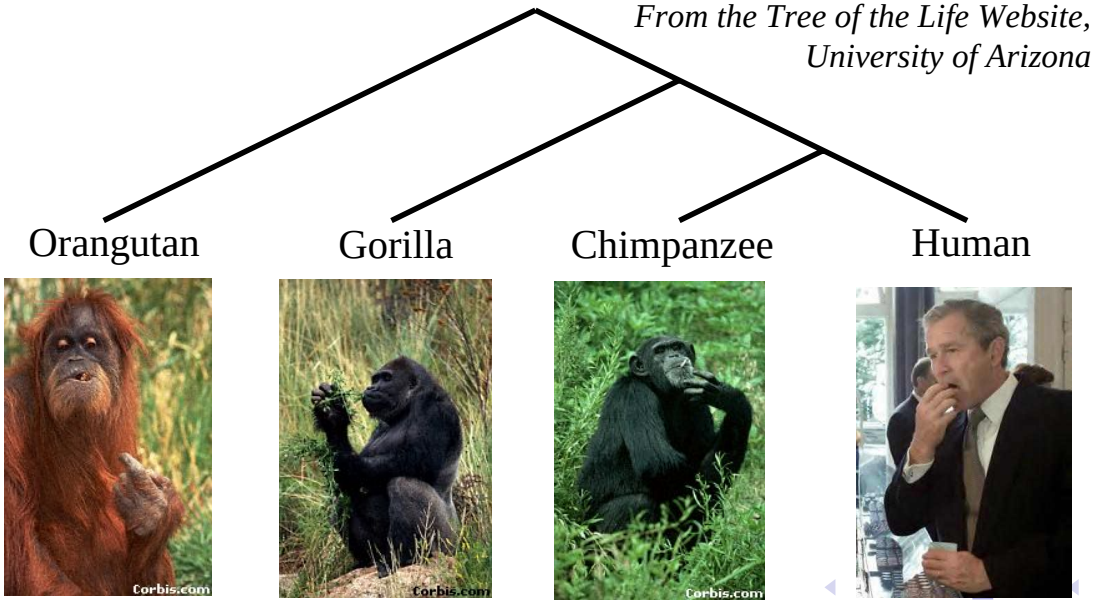


What do we mean when we draw a tree?

- the arrow of time is down the page (or left to right),
- paths down through the tree represent lineages,
- non-leaf vertices represent times at which lineages diverge,
- edge-lengths may have no meaning, or represent chronological time or expected amount of substitution (i.e. mutation),
- the root is the most recent common ancestor of all the taxa.

Species phylogeny

From the Tree of the Life Website, University of Arizona

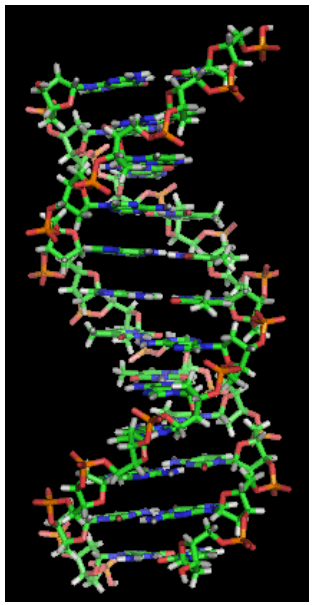


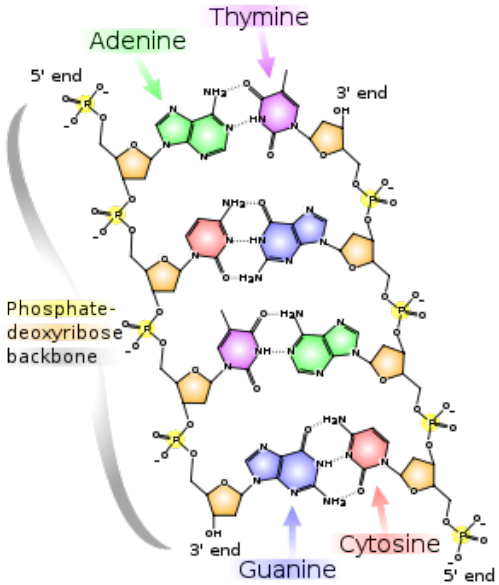
Probability versus statistics

- we describe some aspect of nature with a probability model that has unknown features (parameters)
- statistical inference tries to reconstruct/infer/estimate those parameters from data
- probability theory is engineering
- statistical inference is reverse-engineering
- forward problems versus inverse problems

What do the data look like?

- Phylogenetic reconstruction can be based on different kinds of data: morphology, gene order, etc. I'll concentrate on DNA sequence data.
- For each taxon we have DNA sequences (i.e. strings of nucleotides A=adenine, G=guanine, C=cytosine, T=thymine) for parts of their genome that are "comparable".
- The DNA sequences of the different taxa differ. As evolution occurs, one nucleotide is substituted for another, segments of DNA are deleted, and new segments are inserted.





Sequence alignment

- Sequence alignment procedures are algorithms that:
  - take DNA sequences from several taxa,
  - line up common positions at which substitutions may or may not have occurred,
  - determine where deletions and insertions have occurred in certain sequences relative to the others.
- For example, an alignment of two taxa might produce an output such as the following:

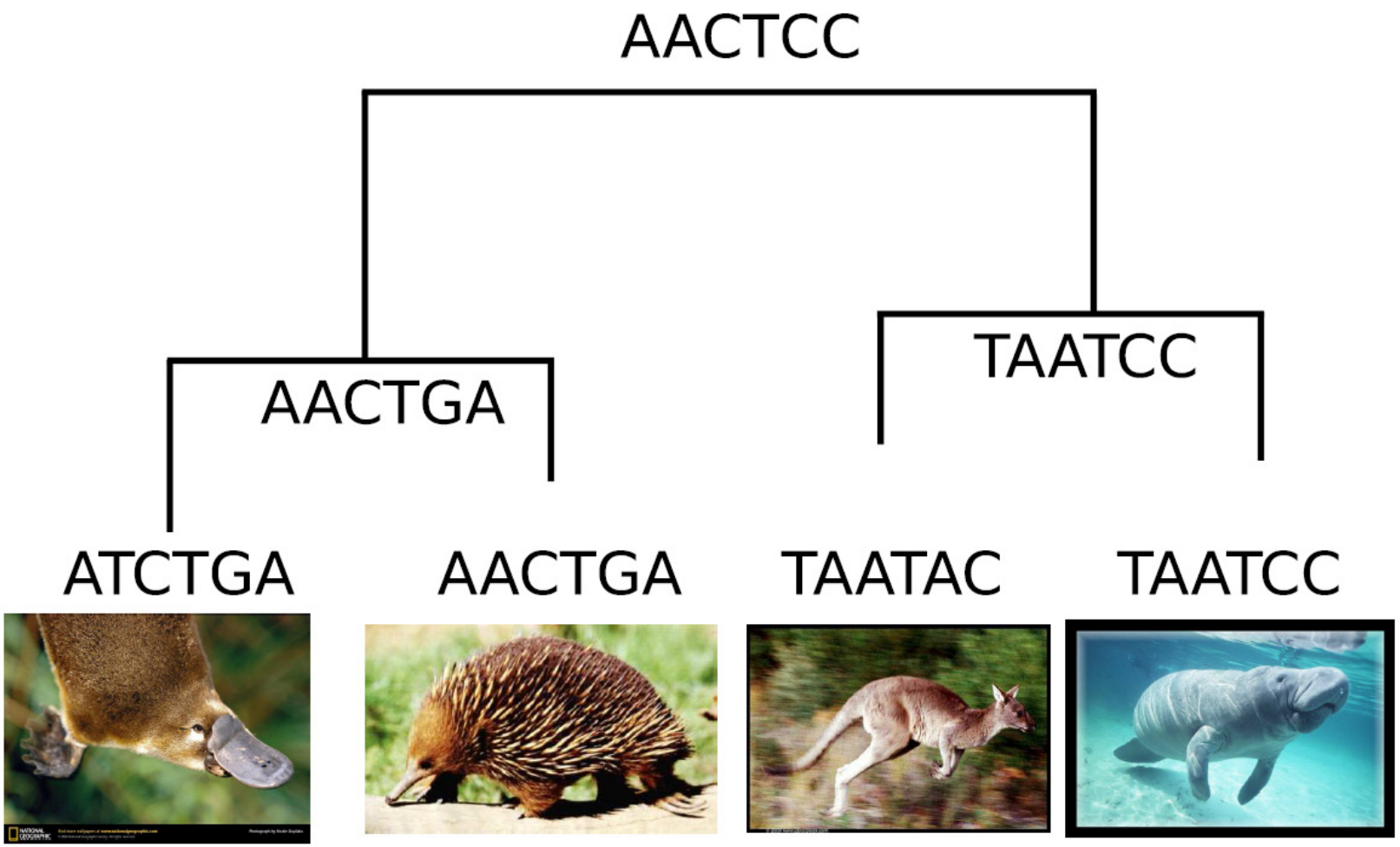
$$\begin{array}{ll} \text{Taxon 1} & \ldots AGTAACT\ldots \\ \text{Taxon 2} & \ldots AT***CA\ldots \end{array}$$

  Reading from left to right:
  - both taxa have an A in the "same" position,
  - the next position is common to both taxa but Taxon 1 has a G there whereas Taxon 2 has a T,
  - then (due to insertions or deletions) there is a stretch of 3 positions that are present in the genome of Taxon 1 but not present in the genome of Taxon 2 *etc.*
- A discussion of alignment algorithms is outside the scope of this course.

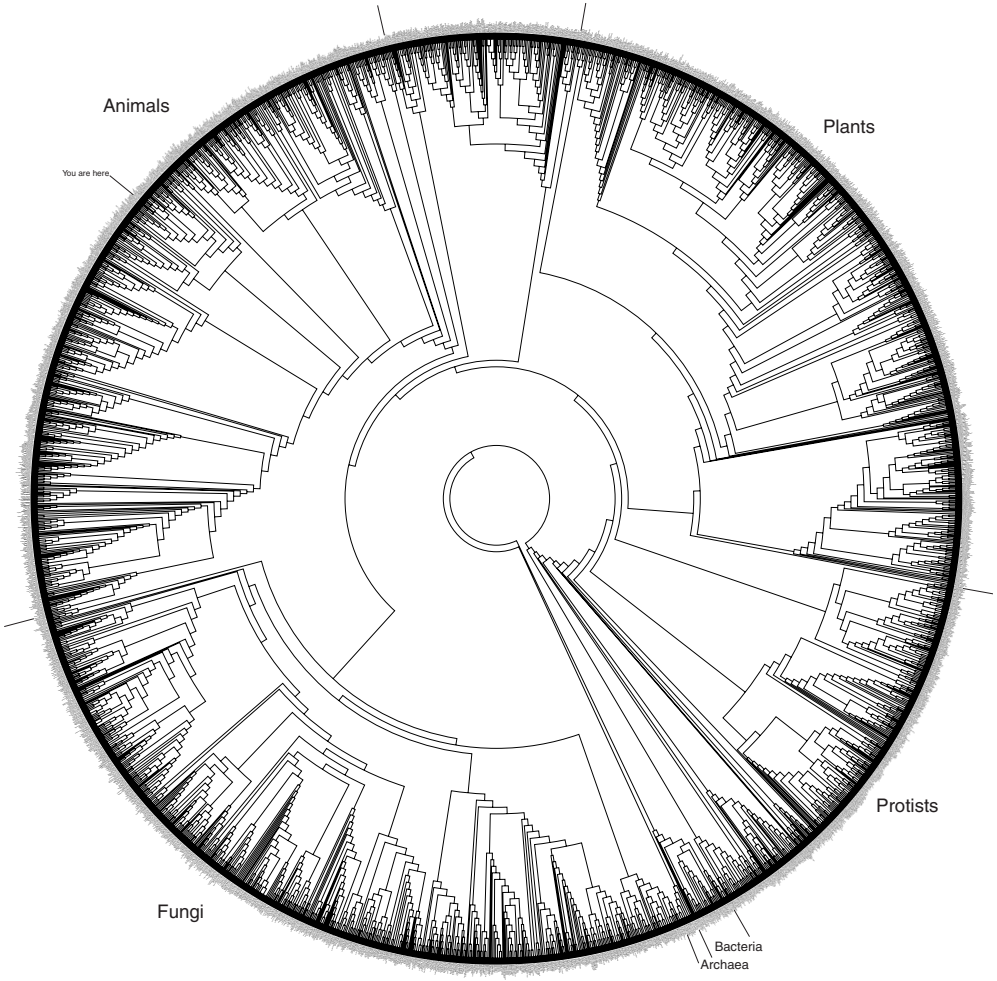
Aligned data

- From now on, suppose that the original sequences have been preprocessed in some suitable way to align them.
- For simplicity, suppose that we are dealing with segments where there have been no insertions or deletions, so all the taxa share the same common positions and differences between nucleotides at these positions are due to substitutions.

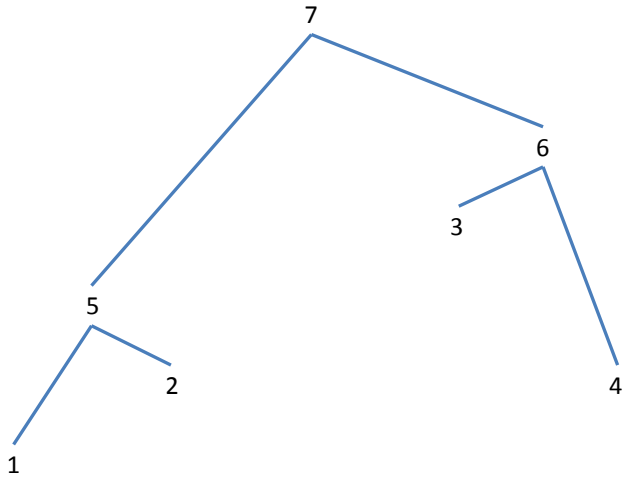
Our goal



Our ultimate goal?

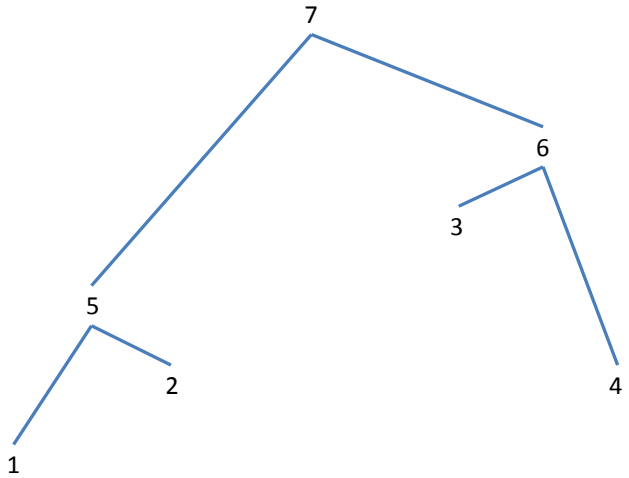


The standard (?) phylogenetic model – Step 1



- Write Y_i for the denote the nucleotide exhibited by "individual" i for some site. (Note: The taxa are $1, 2, 3, 4$ – we do not observe the ancestors $5, 6, 7$.)
- We assume
  - Y_1 and Y_2 are conditionally independent given Y_5,
  - Y_3 and Y_4 are conditionally independent given Y_6,
  - the pair (Y_1, Y_2) is conditionally independent of the pair (Y_3, Y_4) given any one of Y_5, Y_6, or Y_7.

Standard model – Step 1 – continued



Because of this dependence structure, a joint probability such as

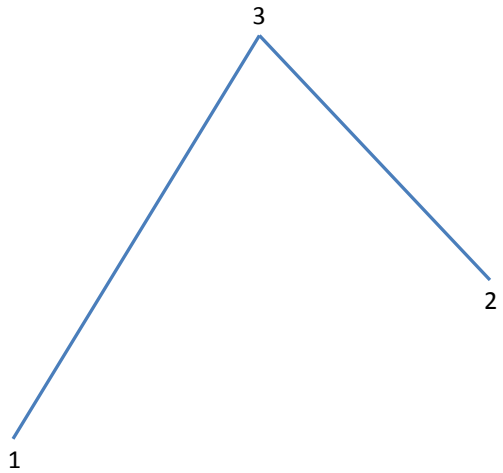
$$\mathbb{P}\{Y_1 = A, Y_2 = A, Y_3 = G, Y_4 = C, Y_5 = T, Y_6 = T, Y_7 = A\}$$

can be computed as

$$\begin{aligned}\mathbb{P}\{Y_7 = A\} \times \mathbb{P}\{Y_5 = T \,|\, Y_7 = A\}\mathbb{P}\{Y_6 = T \,|\, Y_7 = A\} \\ \times \mathbb{P}\{Y_1 = A \,|\, Y_5 = T\}\mathbb{P}\{Y_2 = A \,|\, Y_5 = T\} \\ \times \mathbb{P}\{Y_3 = G \,|\, Y_6 = T\}\mathbb{P}\{Y_4 = C \,|\, Y_6 = T\}.\end{aligned}$$

Standard model – Step 2

To get the probability of the data (that is, the likelihood) for a single site we "sum over" the unobserved "individuals". For example, for the 2 taxon tree



$$
\begin{aligned}
&\mathbb{P}\{Y_1 = A, Y_2 = G\} \\
&\quad = \mathbb{P}\{Y_3 = A\}\mathbb{P}\{Y_1 = A \,|\, Y_3 = A\}\mathbb{P}\{Y_2 = G \,|\, Y_3 = A\} \\
&\qquad + \mathbb{P}\{Y_3 = G\}\mathbb{P}\{Y_1 = A \,|\, Y_3 = G\}\mathbb{P}\{Y_2 = G \,|\, Y_3 = G\} \\
&\qquad + \mathbb{P}\{Y_3 = C\}\mathbb{P}\{Y_1 = A \,|\, Y_3 = C\}\mathbb{P}\{Y_2 = G \,|\, Y_3 = C\} \\
&\qquad + \mathbb{P}\{Y_3 = T\}\mathbb{P}\{Y_1 = A \,|\, Y_3 = T\}\mathbb{P}\{Y_2 = G \,|\, Y_3 = T\}.
\end{aligned}
$$

Standard model – Step 3

- So far, we have a model for what happens at a single site with parameters:
  - the underlying tree
  - $(4-1)=3$ numerical parameters at the root (= distribution of root nucleotide)
  - $4(4-1)$ numerical parameters per edge (= conditional probabilities describing what happens on and edge = substitution probabilities)
- To model what happens jointly at different sites, the most common assumption is that different sites are independent with the same tree, but possibly different root and substitution probabilities $\Longrightarrow$ likelihood for the data is the product of the likelihoods for each site.
- Some models have the substitution probabilities for each $(\text{site}, \text{edge})$ pair as independent picks from a fixed distribution or variants thereof (associated with jargon such as rates-across-sites).

Markov chains models for substitution

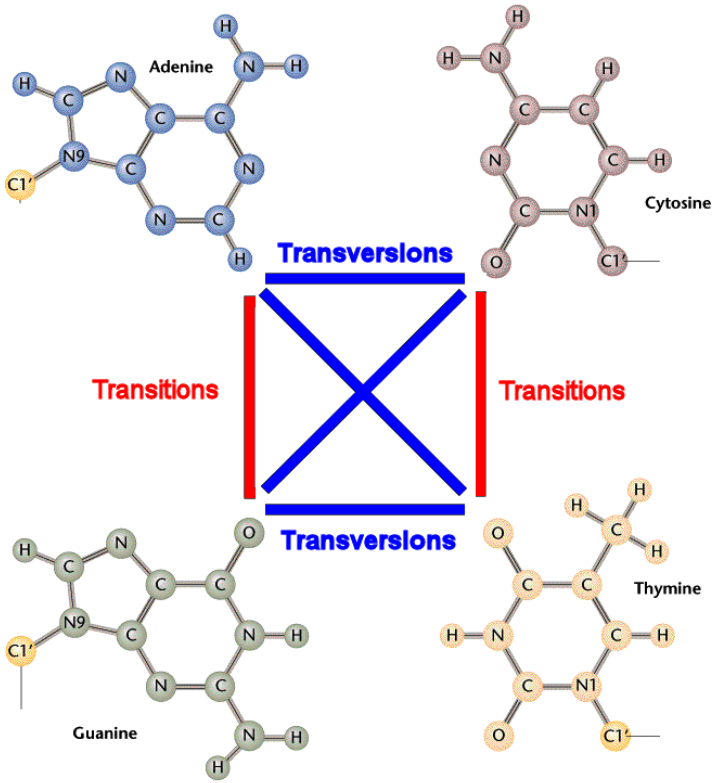
- The matrix of substitution probabilities for an edge captures effect of the substitutions that occurred between the times that the individuals associated with the start and finish of the edge were alive.
- We should model the time dynamics of this substitution process.
- The most natural and tractable dynamics are those of a (time-homogeneous) Markov chain. That is, if the site currently exhibits a certain nucleotide, B' say, then (independently of the past) the nucleotide changes at rate $q(B', B'')$ to some other nucleotide B''.
- More formally, if the position currently exhibits nucleotide B', then:
  - independently of the past, the probability that the elapsed time until a change occurs is greater than t is $\exp(-\sum_{B'' \neq B'} q(B', B'')\, t)$,
  - independently of how long it takes until a change occurs, the probability that it is to B'' is proportional to $q(B', B'')$.

The Jukes-Cantor model

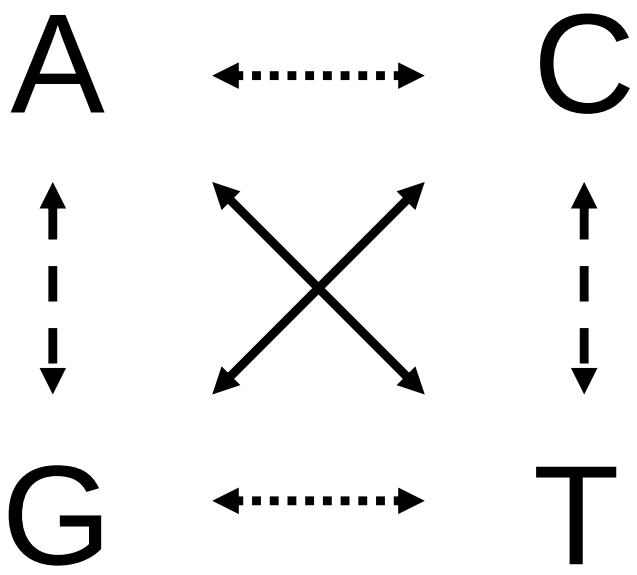
- The simplest Markov chain model for nucleotide substitution is the Jukes-Cantor model in which $q(B', B'')$ is the same for all $B' \neq B''$.
- Under this model,
  - the distribution of the amount of time spent at the current nucleotide before a change occurs does not depend on the nucleotide,
  - all 3 choices of the new nucleotide are equally likely when a change occurs.

Biochemistry of the DNA nucleotides

- Biochemically, the nucleotides fall into two families: the purines (adenine and guanine) and the pyrimidines (cytosine and thymine).
- Substitutions within a family are called transitions, and they have a different biochemical status to substitutions between families, which are called transversions.



Kimura's models



Arrows with of the same type have the same rate.
A sub-model has the same rates for all transversions.

Computing the substitution matrices

- Probabilists usually record the rates for a Markov chain as an infinitesimal generator matrix Q that has $q(B', B'')$ in position (B', B''), $B' \neq B''$, and $q(B', B') = -\sum_{B'' \neq B'} q(B', B'')$, so the rows of Q sum to zero.
- For example, the infinitesimal generator for the 3 parameter Kimura model is

$$Q = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} -(\alpha+\beta+\gamma) & \alpha & \beta & \gamma \\ \alpha & -(\alpha+\beta+\gamma) & \gamma & \beta \\ \beta & \gamma & -(\alpha+\beta+\gamma) & \alpha \\ \gamma & \beta & \alpha & -(\alpha+\beta+\gamma) \end{pmatrix} \end{array}.$$

Computing the substitution matrices - continued

- The infinitesimal generator is more than just an accounting device: for any $s, t \geq 0$ the entry in row B' and column B'' of the matrix

$$\exp(tQ) = I + tQ + \frac{t^2}{2!}Q^2 + \frac{t^3}{3!}Q^3 + \cdots$$

  gives the conditional probability that nucleotide B'' will be exhibited at time $s + t$ given that nucleotide B' is exhibited at time s.
- In particular, if an edge has length u, then the substitution matrix for the edge is $\exp(uQ)$.

Finding estimators

NOW THAT WE HAVE A PROBABILITY MODEL FOR OUR DATA, HOW DO WE USE IT TO ESTIMATE THE CORRECT TREE?

Maximum likelihood

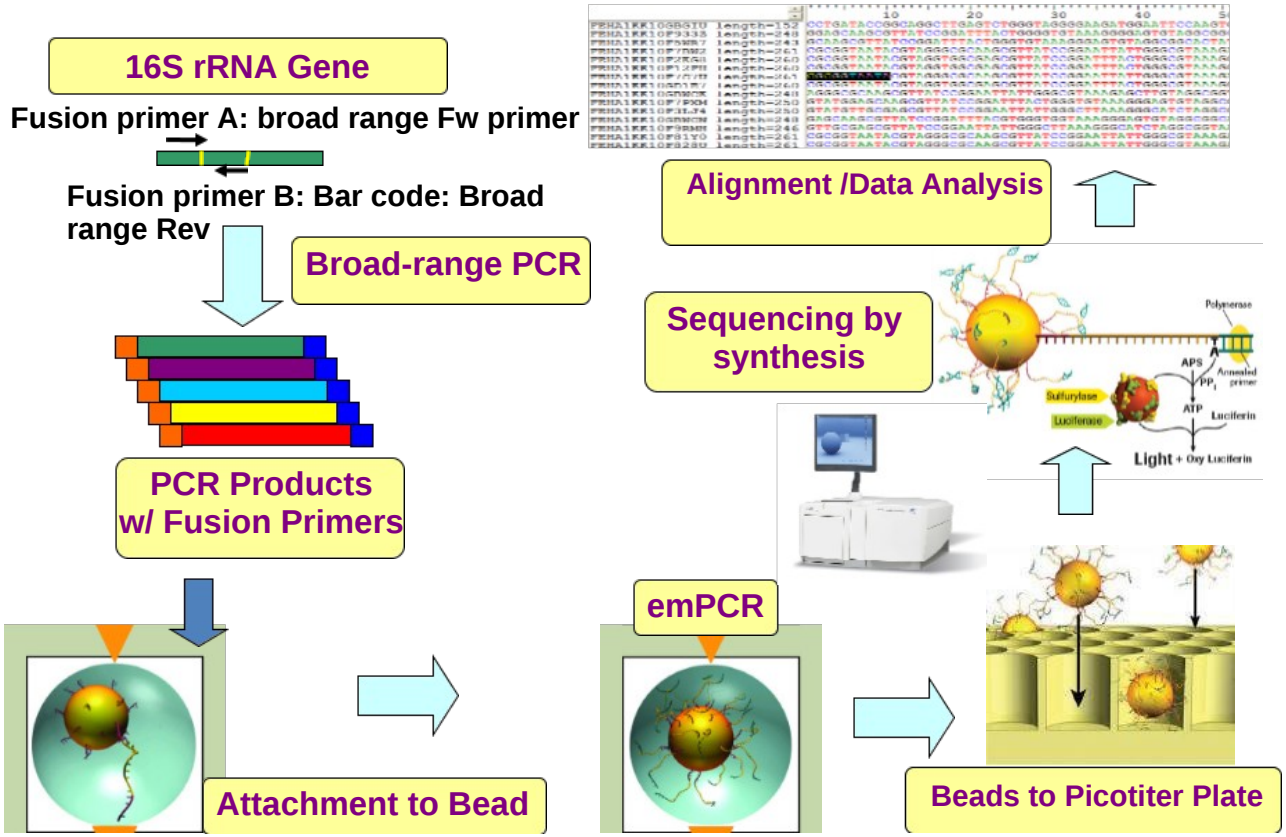
- Estimate the parameters using the choice of parameters that maximizes the likelihood (i.e. makes the data most probable).
- This problem can't be solved analytically in phylogeny. It is usually computationally expensive (high-dimensional optimization), and involves explicitly estimating the numeric parameters even if we only want the shape of the tree.
- Even numerically, heuristics have to be used for large numbers of taxa (e.g. RAxML).

Metagenomics

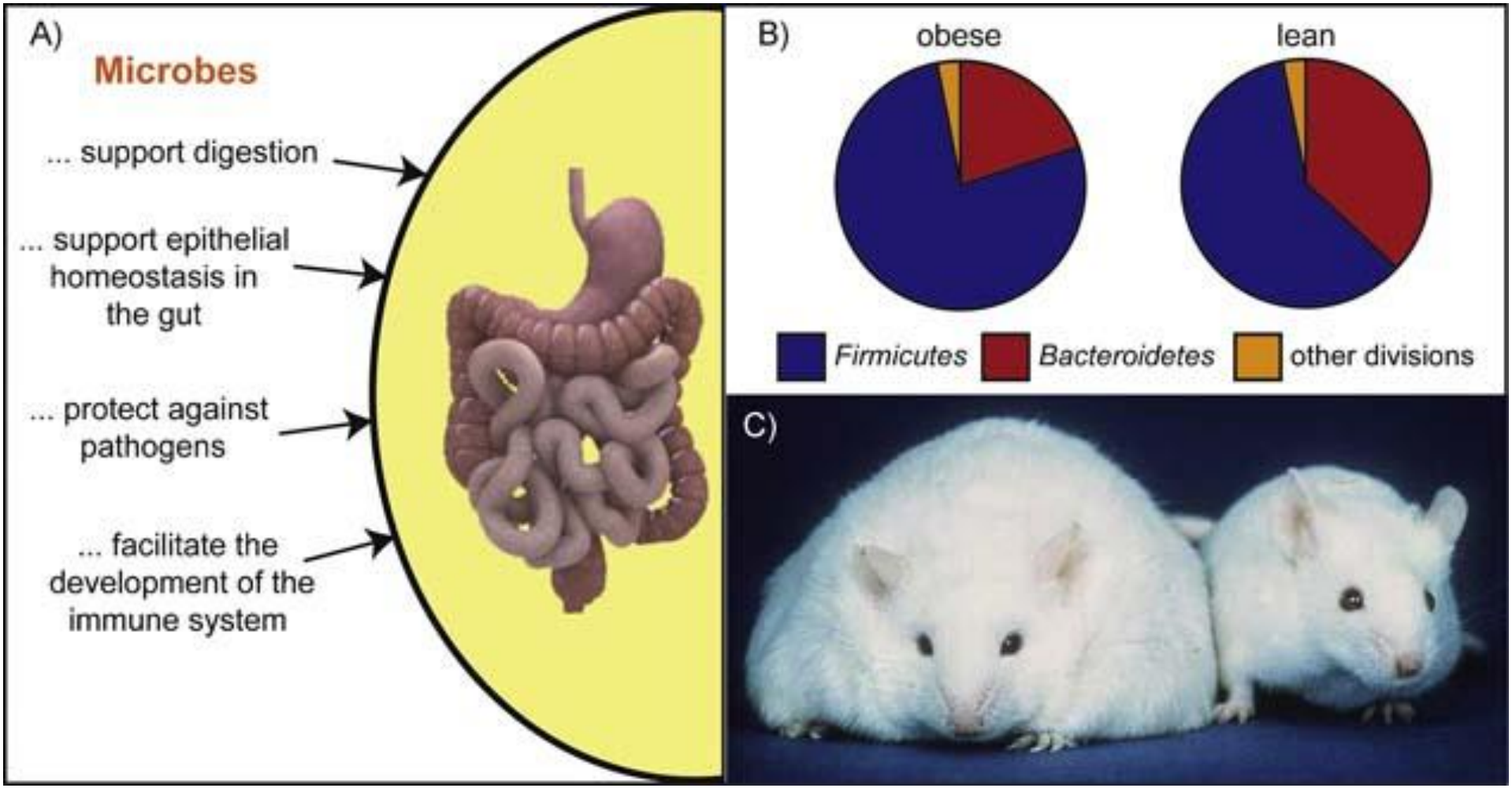
- Next-generation sequencing technology enables sequencing of hundreds of thousands to millions of short DNA sequences in a single experiment.
- Microbial genetic material can be extracted in bulk from a sample taken from some environment and directly sequenced.
- It is no longer necessary to identify individual species by morphology or culturing experiments.
- This technology has revolutionized the possibilities for surveys of environmental microbial diversity, ranging from the human gut (*Gill et al., 2006*) to acid mine drainages (*Baker and Banfield, 2003*).

How is this done?

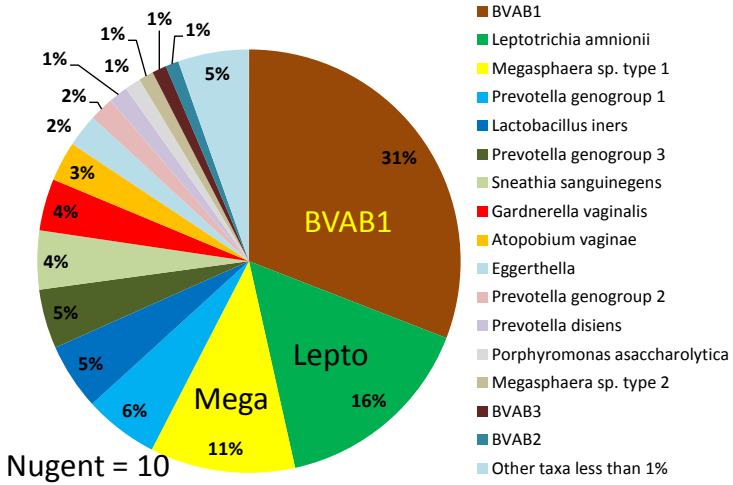
Schematic for Pyrosequencing Approach



Why might a survey of a "microbiome" be useful?



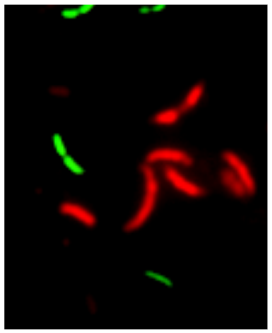
Why might a survey of a "microbiome" be useful? – continued

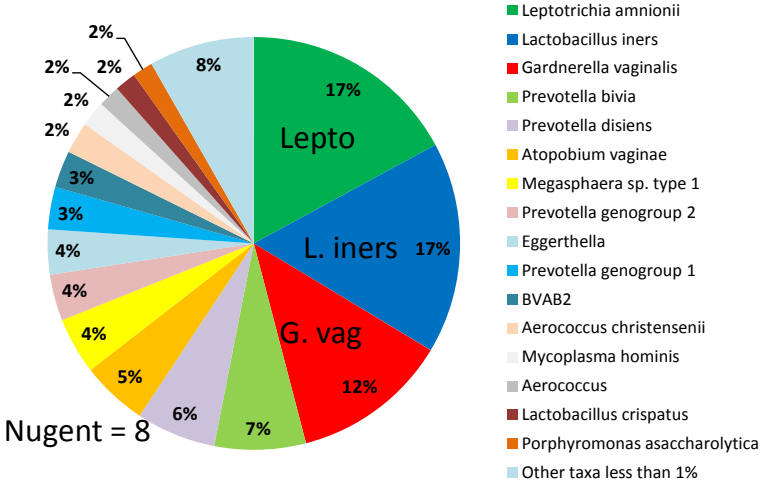


Nugent = 10

Women with BV

- Comparison of Nugent Score 10 and Nugent score 8
- Predominant taxon is BVAB1 when the Nugent Score is 10 – curved morphotype
- *L. iners* often called as *Gardnerella* morphotype
- Quantitative PCR – Median values
  - BVAB1: 5.1 x 10^8 copies
  - *Mobiluncus* spp: 3.7 x 10^5 copies





Nugent = 8

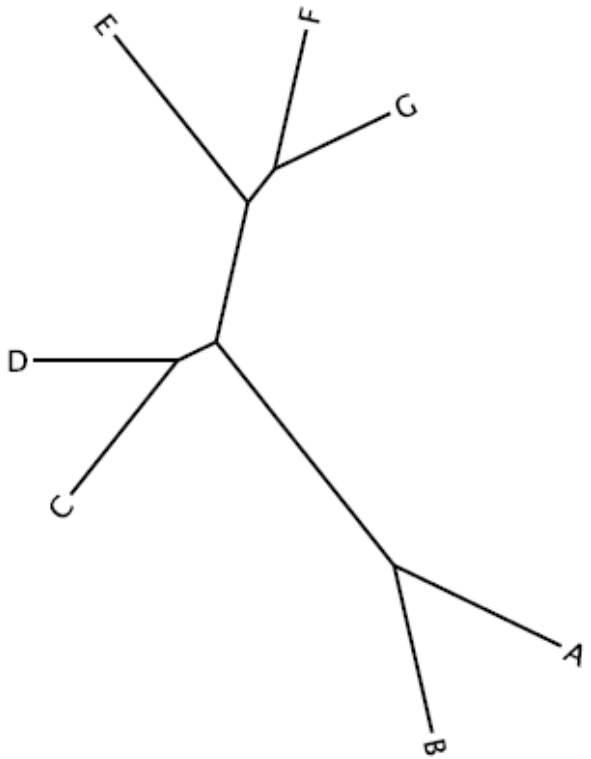
What can we do with this data?

- Huge amount of data – 400K sequences from a single run.
- "Reads" (little shreds of DNA) are often short and non-overlapping.
- Not enough signal in the data to resolve a phylogenetic tree.
- Computationally infeasible to build a tree anyway.

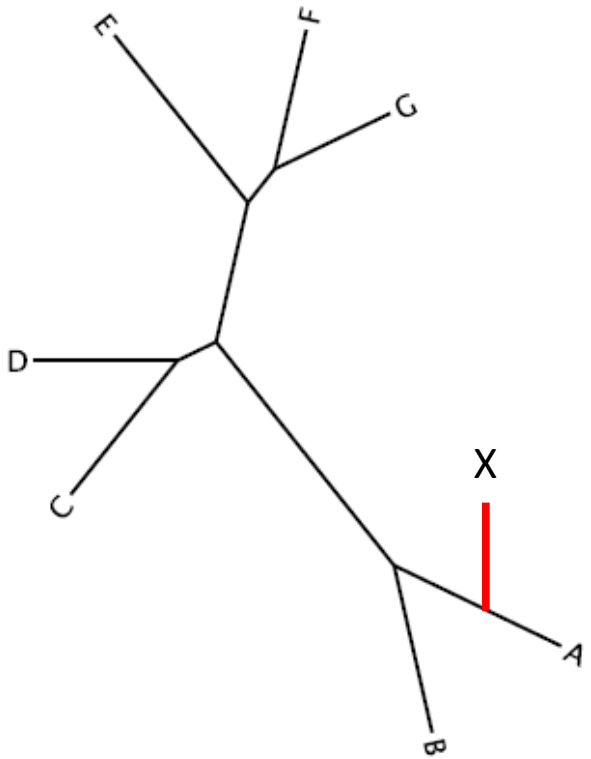
Phylogenetic placement

- Begin with a reference phylogenetic tree constructed from previously-characterized DNA sequences.
- Place query sequences at their most likely position on the reference tree.
- Recent such algorithms are able to place tens of thousands of query sequences on a reference tree of one thousand taxa (species).
- Rather than point placements, likelihood weight ratios can be used to approximate a posterior probability and spread a placement out as a probability distribution on the reference tree.

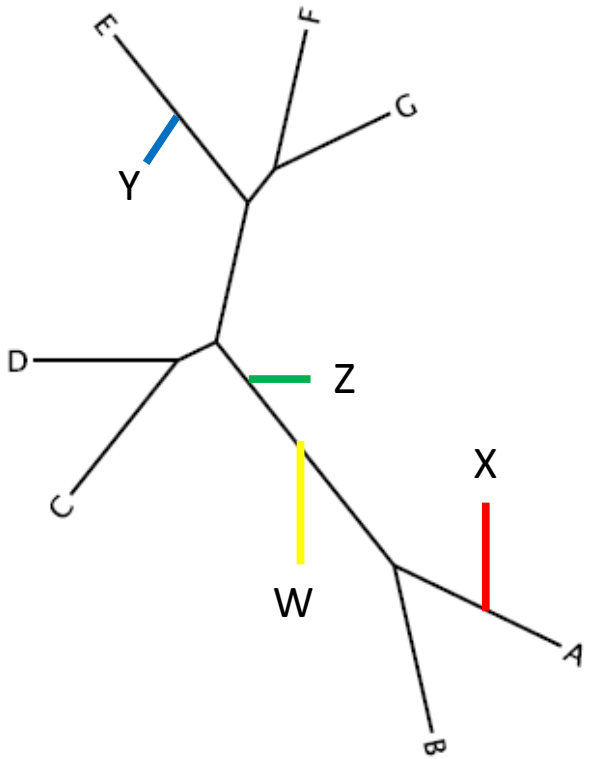
What we start with



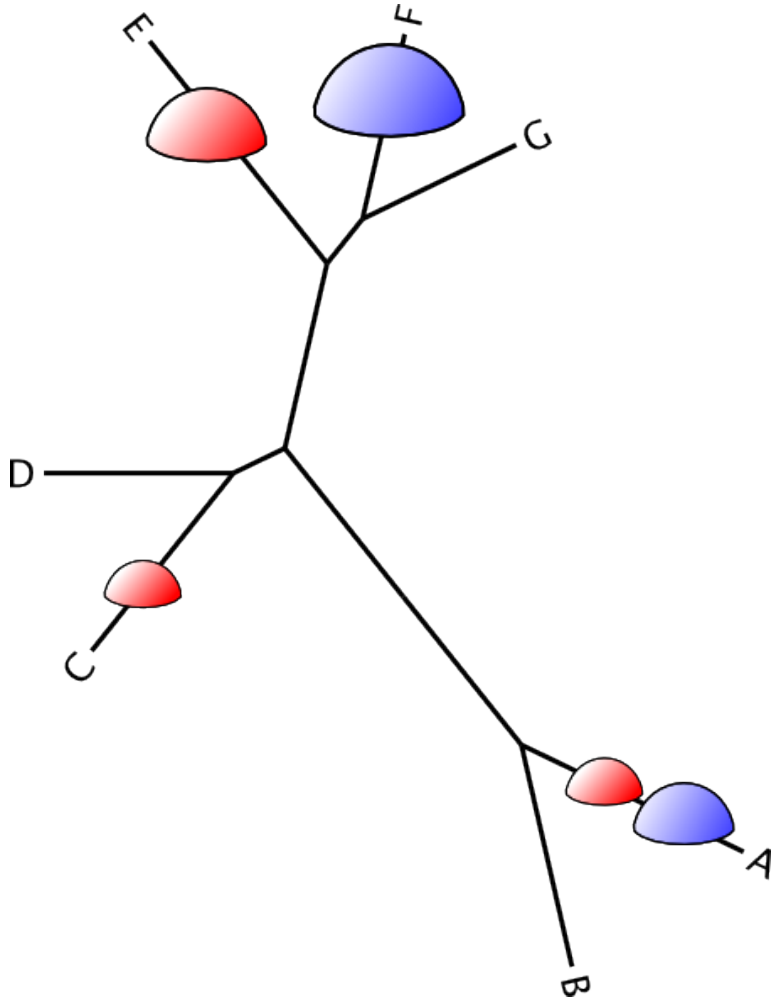
One placement



What we end up with



Spread out placements



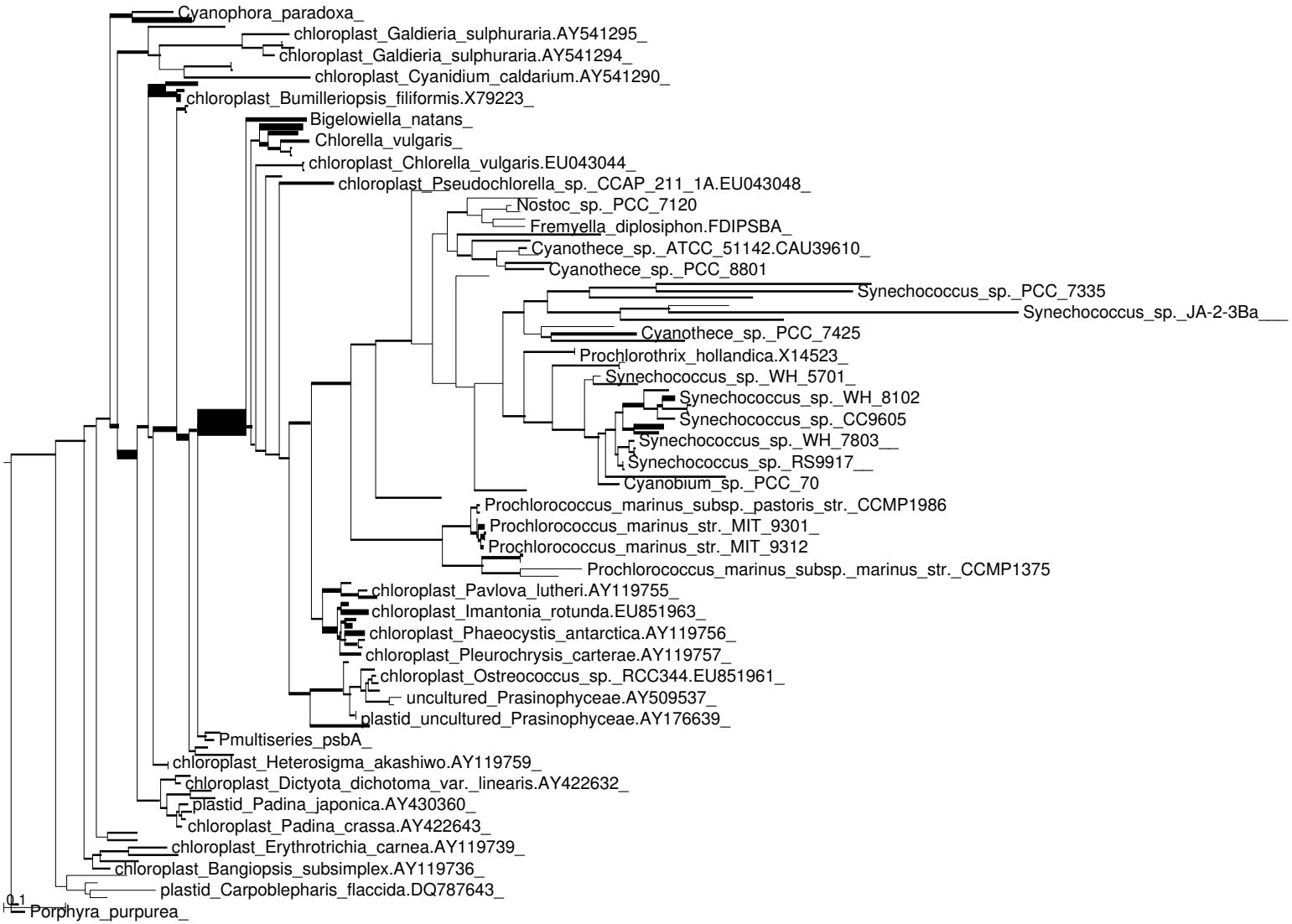
An example

- (*Vila-Costa et al, 2010*): compare communities of marine bacteria in the Sargasso Sea with and without dimethylsulfoniopropionate (DMSP) enrichment.
- Their data was downloaded from the CAMERA website and alignment of the *psbA* gene was supplied by Robin Kodner (UW Friday Harbor Laboratories).
- Searching and alignment was performed using HMMER (*Eddy, 1998*).
- Phylogenetic placement was performed using pplacer (*Matsen, Kodner and Armbrust, 2010*).
- 909 control reads and 451 DMSP treatment reads

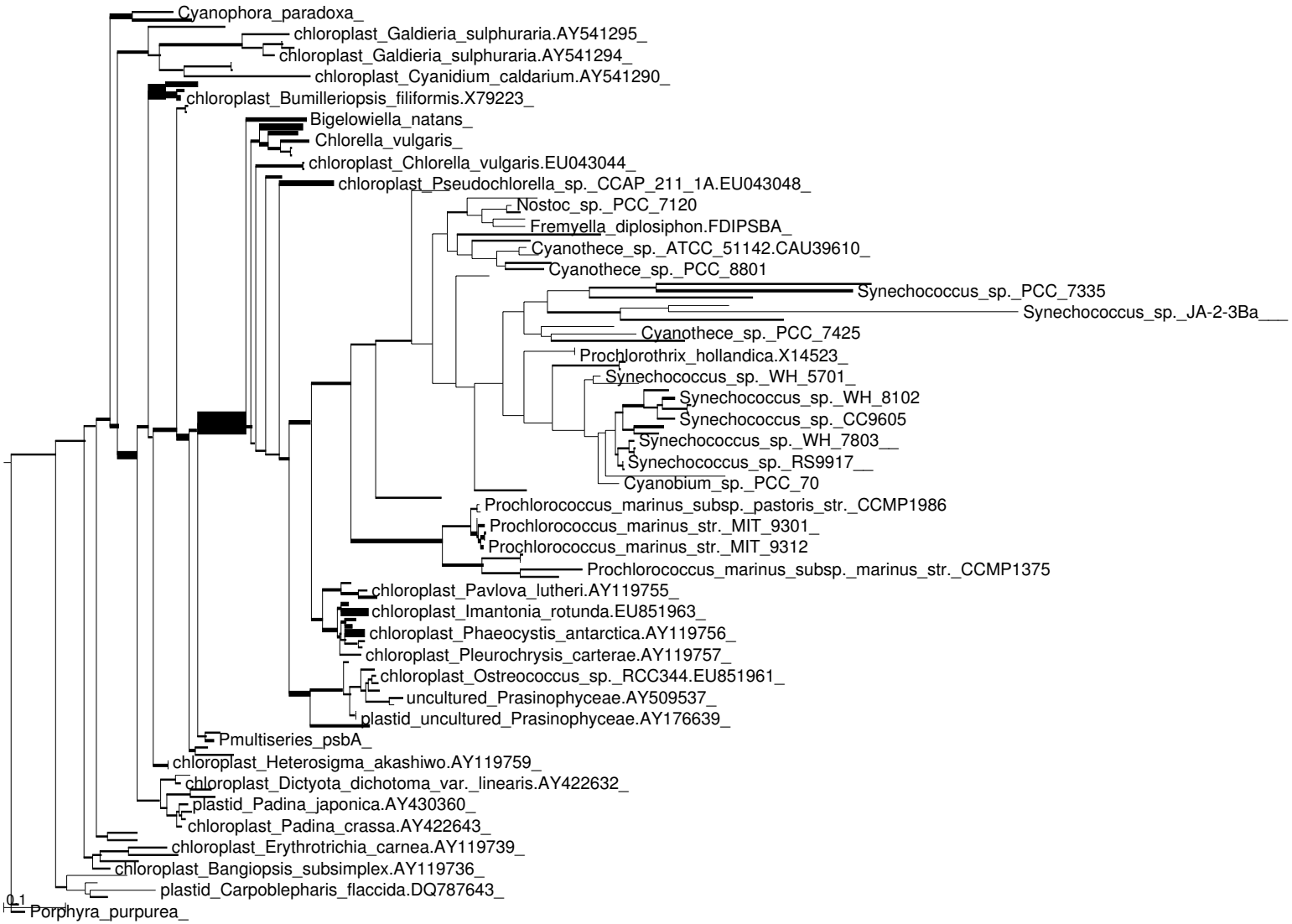
A mathematical framework

- Mathematically, we have m placements in the control sample and n placements in the treatment sample.
- Each control (resp. treatment) placement is a probability measure $\mathbb{P}_i$ (resp. $\mathbb{Q}_j$) on the reference tree $\mathcal{T}$.
- Distinguishing between the two samples is a matter of comparing the probability measures $\mathbb{P} := \frac{1}{m}\sum_i \mathbb{P}_i$ and $\mathbb{Q} = \frac{1}{n}\sum_j \mathbb{Q}_j$.

Edge lengths fattened by control placement weight

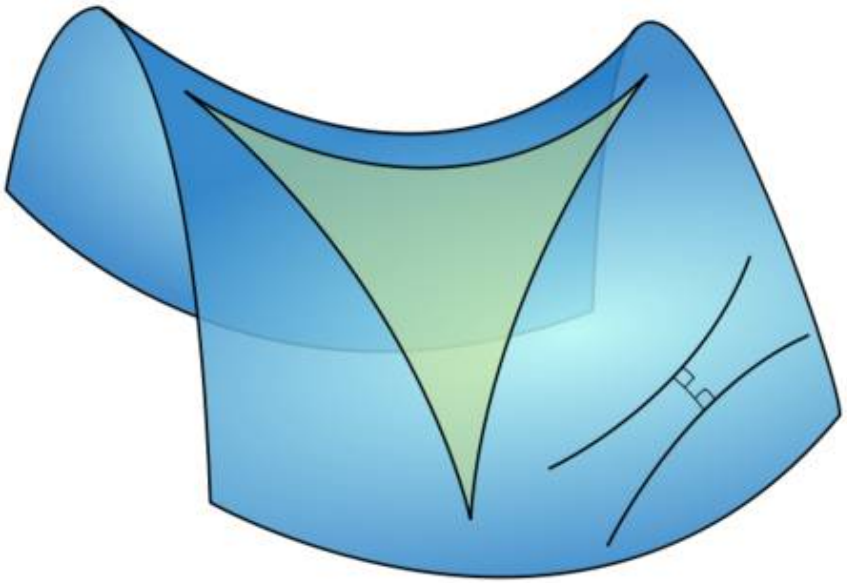


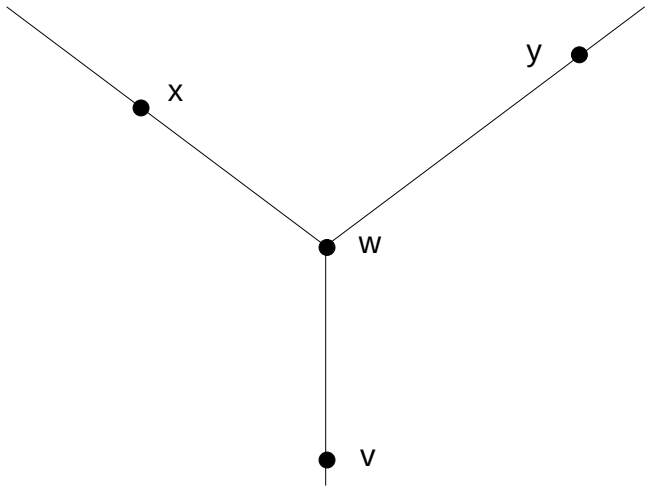
Edge lengths fattened by treatment placement weight



Probability measures on trees have barycenters

- The reference tree $\mathcal{T}$ has a metric d on it (evolutionary distance).
- The metric space $(\mathcal{T}, d)$ is Hadamard (i.e. complete, simply connected and negatively curved in a suitable sense).
- Because it is a Hadamard space with bounded diameter, probability measures on $\mathcal{T}$ have barycenters: if $\mathbb{P}$ is a probability measure on $\mathcal{T}$, then there is a unique $\mu \in \mathcal{T}$ minimizing $x \mapsto \int_{\mathcal{T}} d(x,y)^2 \, \mathbb{P}(dy)$.
- Any point $x \in \mathcal{T}$ splits $\mathcal{T}$ into two or more subtrees $\mathcal{S}$. It is straightforward to calculate μ from a knowledge of $\int_{\mathcal{S}} d(x,y) \, \mathbb{P}(dy)$ for all x and $\mathcal{S}$.





Bare-hands proof that barycenters exist

- As a continuous function on a compact metric space, the function $f : \mathcal{T} \to \mathbb{R}_+$ defined by $f(x) := \int_{\mathcal{T}} d(x,y)^2 \, P(dy)$ achieves its infimum.
- Suppose that the infimum is achieved at two points x' and x''. Define a function $\gamma : [0, d(x', x'')] \to [x', x'']$, where $[x', x''] \subseteq \mathcal{T}$ is the path between x' and x'', by the requirement that $\gamma(t)$ is the unique point in $[x', x'']$ that is distance t from x'.
- Check that the composition $f \circ \gamma$ is strongly convex; that is,

$$(f \circ \gamma)(\alpha r + (1-\alpha)s) < \alpha (f \circ \gamma)(r) + (1-\alpha)(f \circ \gamma)(s)$$

  for $0 < \alpha < 1$ and $r, s \in [0, d(x', x'')]$.
- In particular, $f(\gamma(d(x', x'')/2)) = (f \circ \gamma)(d(x', x'')/2) < (f(x') + f(x''))/2$, contradicting the definitions of x' and x''.

How do we find the barycenter?

- For each point $u \in \mathcal{T}$ there is the associated set of directions in which it is possible to proceed when leaving u: there is one direction for every connected component of $\mathcal{T} \setminus \{u\}$.
- Given a point $u \in \mathcal{T}$ and a direction δ, write $\mathcal{T}(u, \delta)$ for the subset of $\mathcal{T}$ consisting of points $v \neq u$ such that the unique path connecting u and v departs u in the direction δ, set

$$D(u, \delta) := -\int_{\mathcal{T}(u,\delta)} d(u, y)\, \mathbb{P}(dy) + \int_{\mathcal{T} \setminus \mathcal{T}(u,\delta)} d(u, y)\, \mathbb{P}(dy),$$

and note that

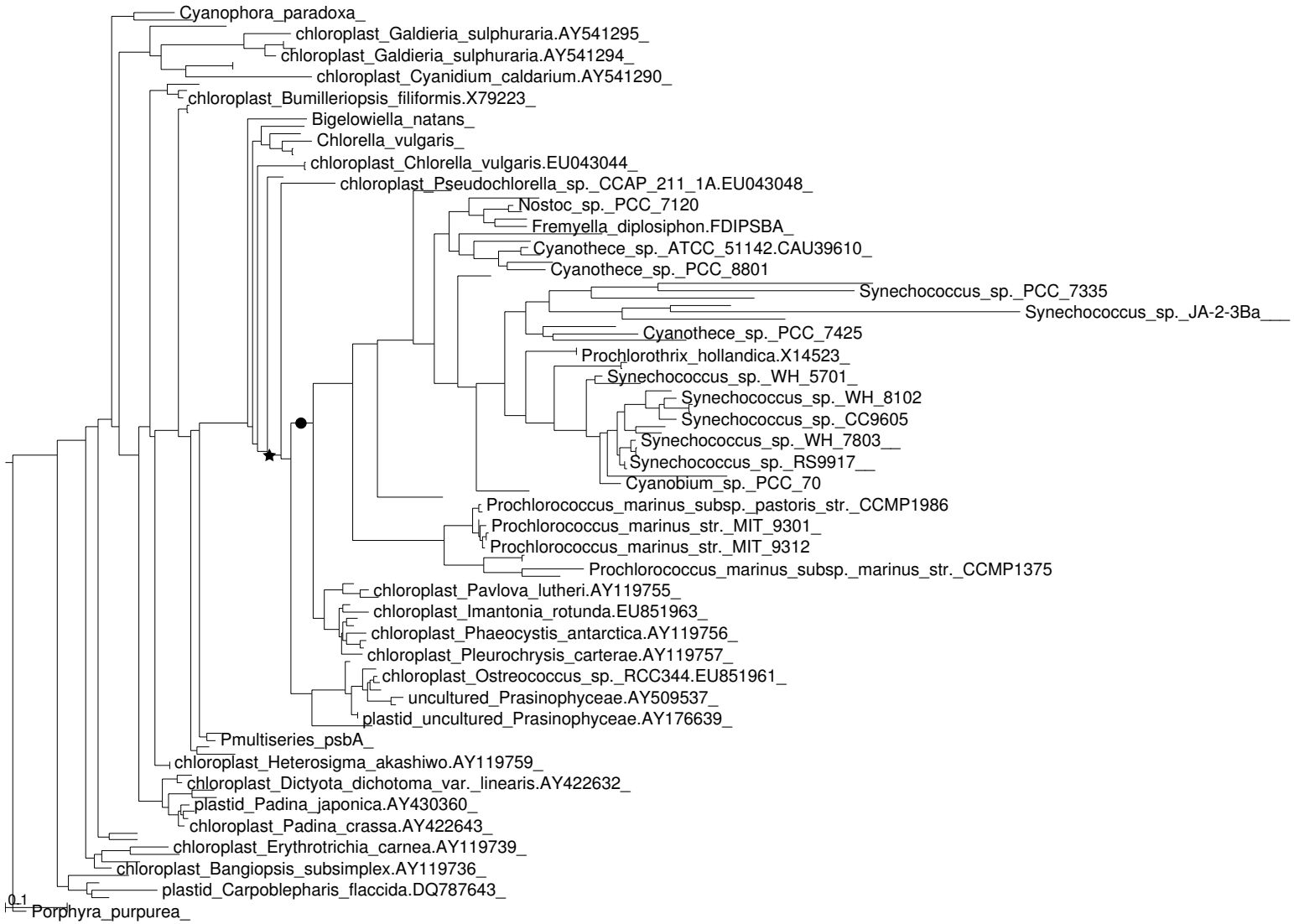
$$\lim_v \frac{1}{d(u, v)} \left[\int_{\mathcal{T}} d(v, y)^2\, \mathbb{P}(dy) - \int_{\mathcal{T}} d(u, y)^2\, \mathbb{P}(dy) \right] = 2D(u, \delta),$$

where the limit is taken over $v \to u$, $v \in \mathcal{T}(u, \delta)$.

How do we find the barycenter? - continued

- If for some vertex u of the reference tree $D(u, \delta) \geq 0$ for all directions δ associated with u, then u is the barycenter (this case includes the trivial one in which u is a leaf and all the mass of $\mathbb{P}$ is concentrated on u).
- If there is no such vertex, then there must be a unique pair of neighboring vertices a and b such that $D(a, \alpha) < 0$ and $D(b, \beta) < 0$, where α is the direction from a pointing towards b and β is the direction from b pointing towards a. In that case, the barycenter must lie on the edge between a and b, and the barycenter is the point $u \in (a, b)$ such that $d(a, u) = -D(a, \alpha)$.

Our two samples have very similar barycenters!



controls $= \bullet$ DSMP $= *$

How far apart are two probability measures? - first attempt

- Suppose that μ and ν are two probability measures on a measurable space $(S, \mathcal{S})$. How do we measure "how far apart" μ and ν are?
- The total variation distance between μ and ν is

$$d_{TV}(\mu, \nu) := \sup_{A \in \mathcal{S}} |\mu(A) - \nu(A)|.$$

- Check that d_{TV} is a metric on the set of probability measures on $(S, \mathcal{S})$.
- Note that $\mu(A) - \nu(A) = \nu(A^c) - \mu(A^c)$, so

$$d_{TV}(\mu, \nu) = \sup_{A \in \mathcal{S}} (\mu(A) - \nu(A)).$$

An explicit formula for the total variation distance

- Choose a measure λ on $(S, \mathcal{S})$ such that $\mu \ll \lambda$ and $\nu \ll \lambda$, with $\mu = \phi \cdot \lambda$ and $\nu = \psi \cdot \lambda$. Then,

$$\mu(A) - \nu(A) = \int_A (\phi(s) - \psi(s))\, \lambda(ds),$$

  and this is clearly maximized by taking $A = \{s \in S : \phi(s) > \psi(s)\}$.

- Consequently,

$$\begin{aligned} d_{TV}(\mu, \nu) &= \int_S (\phi(s) - \psi(s))_+ \, \lambda(ds) \\ &= \int_S (\psi(s) - \phi(s))_+ \, \lambda(ds) \\ &= \frac{1}{2} \int_S |\phi(s) - \psi(s)| \, \lambda(ds) \\ &= \frac{1}{2} \int_S (\phi(s) + \psi(s) - 2\phi(s) \wedge \psi(s)) \, \lambda(ds) \\ &= 1 - \int_S \phi(s) \wedge \psi(s) \, \lambda(ds). \end{aligned}$$

Function formulation of the total variation distance

- Let F be the set of measurable functions $f : S \to \mathbb{R}$ such that $-1 \leq f \leq 1$.
- Check that

$$d_{TV}(\mu, \nu) = \sup_{f \in F} \frac{1}{2} \left(\int_S f(s)\, \mu(ds) - \int_S f(s)\, \nu(ds) \right).$$

Dual description of the total variation distance – coupling

- Suppose that (X, Y) is an $S \times S$-valued random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that X has distribution μ and Y has distribution ν.
- Check that

$$\begin{aligned}\mu(A) - \nu(A) &= \mathbb{P}\{X \in A\} - \mathbb{P}\{Y \in A\} \\ &= \mathbb{E}[\mathbf{1}_A(X) - \mathbf{1}_A(Y)] \\ &\leq \mathbb{E}[\mathbf{1}\{X \neq Y\}] \\ &= \mathbb{P}\{X \neq Y\}.\end{aligned}$$

- Thus,

$$d_{TV}(\mu, \nu) \leq \mathbb{P}\{X \neq Y\}.$$

Dual description of TV and coupling – continued

- Conversely, let I, U, V, W be independent random variables where I is $\{0, 1\}$-valued with

$$\mathbb{P}\{I = 1\} = p := \int_S \phi(s) \wedge \psi(s)\, \lambda(ds)$$

and U, V, W are S-valued with

$$\begin{aligned}
\mathbb{P}\{U \in ds\} &= \phi(s) \wedge \psi(s)\, \lambda(ds)/p, \\
\mathbb{P}\{V \in ds\} &= (\phi(s) - \psi(s))_+\, \lambda(ds)/(1-p), \\
\mathbb{P}\{W \in ds\} &= (\psi(s) - \phi(s))_+\, \lambda(ds)/(1-p).
\end{aligned}$$

Set

$$(X, Y) = \begin{cases} (U, U), & \text{if } I = 1, \\ (V, W), & \text{if } I = 0. \end{cases}$$

- Check that X has distribution μ, Y has distribution ν, and

$$\mathbb{P}\{X \neq Y\} = (1 - p) = d_{TV}(\mu, \nu).$$

Dual description of TV and coupling – continued

Therefore, if we write R for the set of probability measures ρ on $(S \times S, \mathcal{S} \times \mathcal{S})$ with the property $\rho(A \times S) = \mu(A)$ and $\rho(S \times B) = \nu(B)$, then

$$d_{TV}(\mu, \nu) = \inf_{\rho \in R} \int_{S \times S} r(x, y)\, \rho(dx, dy),$$

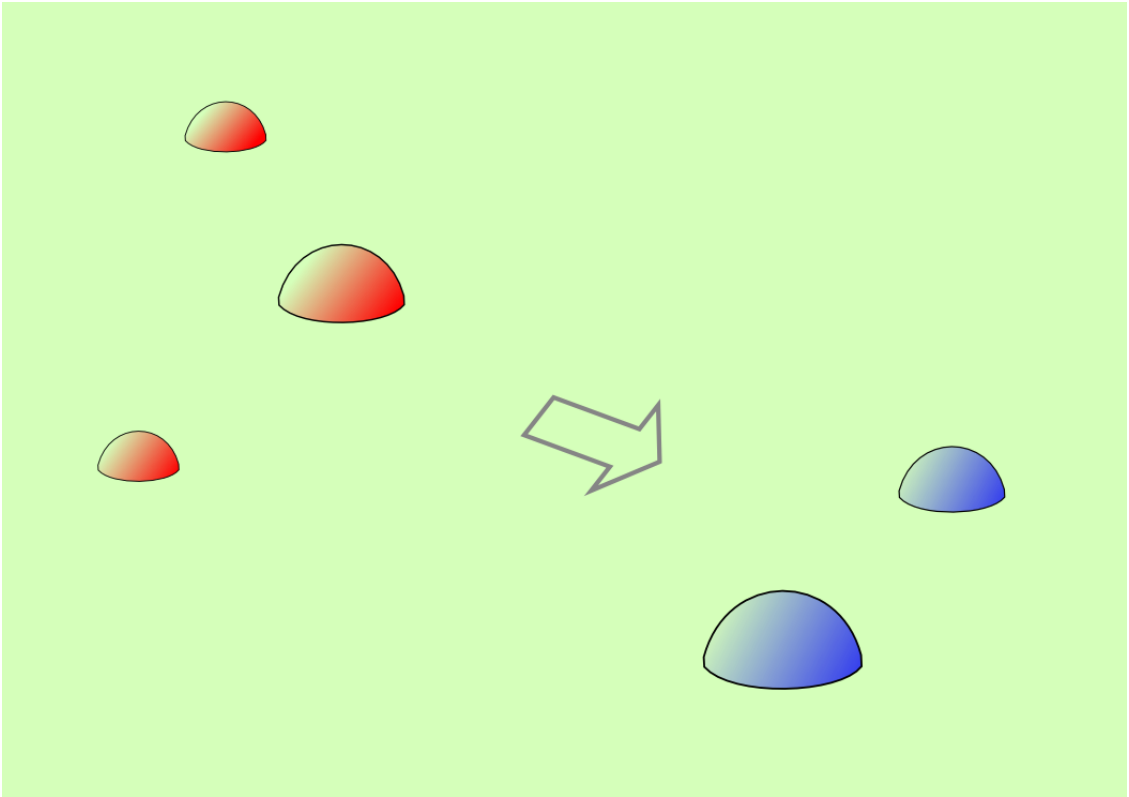
where r is the discrete metric on S given by

$$r(x, y) = \begin{cases} 0, & \text{if } x = y, \\ 1, & \text{otherwise.} \end{cases}$$

The total variation distance isn't that useful

- The total variation distance pays no attention to the evolutionary distance structure on the tree: if one took k point placements and constructed another set of placements by perturbing each of the original placements by a tiny amount so that the two sets of placements were disjoint, then the total variation distance between the corresponding probability distributions would be 1, the largest it can be for any pair of probability distributions, even though we should regard the two sets of placements as being very close. (Even genetic material from organisms of the same species can result in slightly different placements due to genetic variation within species and experimental error.)
- We need a metric that incorporates the evolutionary distance on the reference tree and measures two distributions as being close if one is obtained from the other by short range redistributions of mass.

Probability measures as piles of dirt



Can we compare two probability measures by thinking of them as two collections of dirt piles and asking how much work we have to do to transform one into the other?

Gaspard Monge and optimal transport (1781)

MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.

Par M. MONGE.

LORSQU'ON doit transporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit transporter, & le nom de *Remblai* à l'espace qu'elles doivent occuper après le transport.

Le prix du transport d'une molécule étant, toutes choses d'ailleurs égales, proportionnel à son poids & à l'espace qu'on lui fait parcourir, & par conséquent le prix du transport total devant être proportionnel à la somme des produits des molécules multipliées chacune par l'espace parcouru, il s'ensuit que le déblai & le remblai étant donnés de figure & de position, il n'est pas indifférent que telle molécule du déblai soit transportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine distribution à faire des molécules du premier dans le second, d'après laquelle la somme de ces produits sera la moindre possible, & le prix du transport total sera un *minimum*.

C'est la solution de cette question que je me propose de donner ici. Je diviserai ce Mémoire en deux parties, dans la première je supposerai que les déblais & les remblais sont des aires contenues dans un même plan : dans le second, je supposerai que ce sont des volumes.

PREMIÈRE PARTIE.

Du transport des aires planes sur des aires comprises dans un même plan.

I.

QUELLE que soit la route que doive suivre une molécule

déblai = material dug from a quarry
remblai = material put into new construction

Kantorovich-Rubinstein distance

The Kantorovich-Rubinstein distance between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the tree $\mathcal{T}$ with metric d is

$$Z_1(\mathbb{P},\mathbb{Q}) := \inf\left\{\int_{\mathcal{T}\times\mathcal{T}} d(x,y)\,\mathbb{R}(dx,dy)\right\},$$

where the infimum is over all probability measures $\mathbb{R}$ with $\mathbb{R}(A\times\mathcal{T}) = \mathbb{P}(A)$ and $\mathbb{R}(\mathcal{T}\times B) = \mathbb{Q}(B)$.

That is,

$$Z_1(\mathbb{P},\mathbb{Q}) := \inf\left\{\mathbb{E}[d(X,Y)]\right\},$$

where the infimum is over pairs (X,Y) of $\mathcal{T}$-valued random variables such that X has distribution $\mathbb{P}$ and Y has distribution $\mathbb{Q}$.

A dual formulation of KR distance

By convex duality (essentially an infinite dimensional version of the fundamental theorem of linear programming),

$$Z_1(\mathbb{P}, \mathbb{Q}) = \sup \left\{ \int_{\mathcal{T}} f(x)\, \mathbb{P}(dx) - \int_{\mathcal{T}} f(y)\, \mathbb{Q}(dy) \right\},$$

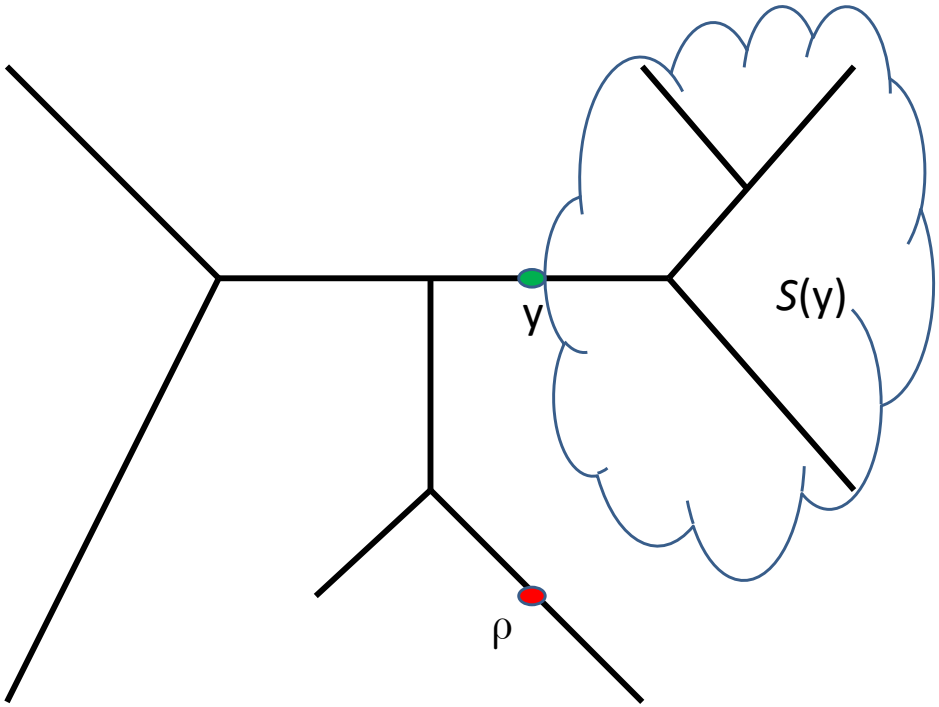
where the supremum is over all functions f with $|f(u) - f(v)| \leq d(u, v)$.

Note the similarity with the dual descriptions of the total variation distance.

WE WILL PROVE THIS LATER.

An explicit formula Kantorovich-Rubinstein distance

Fix a point $\rho \in \mathcal{T}$ (the "root"). For $y \in \mathcal{T}$, let $\mathcal{S}(y)$ be the "subtree on the other side of y".



An explicit formula – continued

Observe that if $h : \mathcal{T} \to \mathbb{R}$ is a bounded Borel function and μ is a Borel probability distribution on $\mathcal{T}$, then we have the integration-by-parts formula

$$\begin{aligned}
\int_{\mathcal{T}} \left(\int_{[\rho,x]} h(y)\, \lambda(dy) \right) \mu(dx) &= \int_{\mathcal{T}\times\mathcal{T}} 1_{[\rho,x]}(y) h(y)\, (\mu \otimes \lambda)(dx, dy) \\
&= \int_{\mathcal{T}\times\mathcal{T}} 1_{\mathcal{S}(y)}(x) h(y)\, (\mu \otimes \lambda)(dx, dy) \\
&= \int_{\mathcal{T}} h(y) \left(\int_{\mathcal{S}(y)} \mu(dx) \right) \lambda(dy) \\
&= \int_{\mathcal{T}} h(y) \mu(\mathcal{S}(y))\, \lambda(dy),
\end{aligned}$$

where λ is the length measure on $\mathcal{T}$.

An explicit formula – continued

- Any function $f : \mathcal{T} \to \mathbb{R}$ with $|f(u) - f(v)| \leq d(u, v)$ can be written as

$$f(x) = C + \int_{[\rho, x]} g(y)\, \lambda(dy)$$

  for some constant C and Borel function $g : \mathcal{T} \to \mathbb{R}$ with $-1 \leq g \leq +1$.
- Thus, if $\mathbb{P}$ and $\mathbb{Q}$ are two probability distributions on $\mathcal{T}$ we have

$$Z_1(\mathbb{P}, \mathbb{Q}) = \sup \left\{ \int_{\mathcal{T}} g(y) \left[\mathbb{P}(\mathcal{S}(y)) - \mathbb{Q}(\mathcal{S}(y)) \right] \lambda(dy) : -1 \leq g \leq +1 \right\}.$$

- It is clear that the integral is maximized by taking $g(y) = +1$ (resp. $g(y) = -1$) when $\mathbb{P}(\mathcal{S}(y)) > \mathbb{Q}(\mathcal{S}(y))$ (resp. $\mathbb{P}(\mathcal{S}(y)) < \mathbb{Q}(\mathcal{S}(y))$), so that

$$Z_1(\mathbb{P}, \mathbb{Q}) = \int_{\mathcal{T}} |\mathbb{P}(\mathcal{S}(y)) - \mathbb{Q}(\mathcal{S}(y))|\, \lambda(dy).$$

Kantorovich-Rubinstein / Wasserstein / Monge-Wasserstein distance

- Consider a compact metric space (S, d).
- For any two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on S, the *Kantorovich-Rubinstein* or *Wasserstein* or *Monge-Wasserstein distance* between them is

$$W(\mathbb{P}, \mathbb{Q}) := \inf \left\{ \int d(x, y)\, d\mu(x, y) : \mu \in M(\mathbb{P}, \mathbb{Q}) \right\},$$

  where $M(\mathbb{P}, \mathbb{Q})$ is the set of all probability measures on $S \times S$ with marginals $\mathbb{P}$ and $\mathbb{Q}$.
- The distance Z_1 for probability measures on a tree $(\mathcal{T}, d)$ is a special case.

The discrete transportation problem

- There are m warehouses with a supply of a certain product, and n shops to which this product must be shipped.
- The i^{th} warehouse possesses an amount p_i.
- The j^{th} shop must receive an amount q_j.
- Suppose that $\sum_i p_i = \sum_j q_j = 1$.
- Let c_{ij} be the cost of shipping one unit from warehouse i to shop j.
- Write $x_{ij} \geq 0$ for the quantity shipped from warehouse i to shop j.

The problem is to minimize the total shipping cost

$$\sum_{ij} x_{ij} c_{ij}$$

subject to the constraints

$$\sum_j x_{ij} = p_i, \quad 1 \leq i \leq m,$$

and

$$\sum_i x_{ij} = q_j, \quad 1 \leq j \leq n.$$

The discrete transportation problem – continued

The dual of the transportation problem involves variables $u_i \geq 0$, $1 \leq i \leq m$, and $v_j \geq 0$, $1 \leq j \leq n$.
The dual problem is to maximize

$$\sum_{j=1}^{n} v_j q_j - \sum_{i=1}^{m} u_i p_i$$

subject to the constraints

$$c_{ij} \geq v_j - u_i, \quad 1 \leq i \leq m,\ 1 \leq j \leq n.$$

The dual may be interpreted as the problem faced by a shipper who comes to the company that owns the warehouses and shops and offers to take care of the company's shipping needs by buying units of the product from warehouse i at price u_i and selling them to shop j at price v_j. The constraint $c_{ij} \geq v_j - u_i$ says that for warehouse i and shop j it is cheaper for the company to use the shipper instead of its own means of transporting the product.

The discrete transportation problem – continued

The fundamental theorem of linear programming says that

$$\min \sum_{ij} x_{ij} c_{ij}$$

subject to

$$x_{ij} \geq 0, \ \sum_j x_{ij} = p_i, \ \sum_i x_{ij} = q_j, \quad 1 \leq i \leq m, \ 1 \leq j \leq n.$$

$$=$$

$$\max \left(\sum_{j=1}^{n} v_j q_j - \sum_{i=1}^{m} u_i p_i \right)$$

subject to

$$u_i \geq 0, \ v_j \geq 0, \ c_{ij} \geq v_j - u_i, \quad 1 \leq i \leq m, \ 1 \leq j \leq n.$$

Lipschitz functions and the dual metric

- The *Lipschitz seminorm* for a real-valued function f on the compact metric space S is

$$||f||_L := \sup\{|f(x) - f(y)|/d(x,y) : x \neq y \in S\}.$$

- For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on S set

$$\begin{aligned}\gamma(\mathbb{P},\mathbb{Q}) &:= \sup\left\{\int f\, d(\mathbb{P}-\mathbb{Q}) : ||f||_L \leq 1\right\} \\ &= \left\{\int f\, d(\mathbb{P}-\mathbb{Q}) : |f(x)-f(y)| \leq d(x,y),\ x,y \in S\right\}.\end{aligned}$$

Kantorovich-Rubinstein theorem

Theorem 1 (Kantorovich-Rubinstein)

For any two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a compact metric space (S, d),

$$W(\mathbb{P}, \mathbb{Q}) = \gamma(\mathbb{P}, \mathbb{Q}),$$

and $W = \gamma$ is a metric on the space of probability measures on (S, d)

Proof. It is clear that γ is a pseudometric and that $W(\mathbb{P}, \mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$, so it suffices to show that $W = \gamma$.

Suppose $\mu \in M(\mathbb{P}, \mathbb{Q})$ and $||f||_L \leq 1$. Then,

$$\int f \, d(\mathbb{P} - \mathbb{Q}) = \int (f(x) - f(y)) \, d\mu(x, y) \leq \int d(x, y) \, d\mu(x, y).$$

Taking the supremum over f, and the infimum over μ, gives

$$\gamma(\mathbb{P}, \mathbb{Q}) \leq W(\mathbb{P}, \mathbb{Q}).$$

Kantorovich-Rubinstein theorem – proof

For the reverse inequality we need some further lemmas.
For any continuous real-valued function h on $S \times S$ and probability measures $\mathbb{P}$ and $\mathbb{Q}$ on S let

$$m_{\mathbb{P},\mathbb{Q}}(h) := \sup \left\{ \int f \, d\mathbb{P} + \int g \, d\mathbb{Q} : f(x) + g(y) < h(x,y) \text{ for all } x, y \right\}.$$

Let $C(S \times S)$ be the space of all continuous real-valued functions on $S \times S$.

Lemma 2

For any compact metric space (S, d), probability measures $\mathbb{P}$ and $\mathbb{Q}$ on S, and $h \in C(S \times S)$,

$$m_{\mathbb{P},\mathbb{Q}}(h) = \inf \left\{ \int h \, d\mu : \mu \in M(\mathbb{P}, \mathbb{Q}) \right\}.$$

Proof. For any $\mu \in M(\mathbb{P}, \mathbb{Q})$, clearly $m_{\mathbb{P},\mathbb{Q}}(h) \leq \int h \, d\mu$. So, "$\leq$" holds in the claimed equation.

For the converse inequality, let L be the linear subspace of functions of the form $\phi(x, y) = f(x) + g(y)$ for continuous functions f and g on S, and set

$$r(\phi) = \int f \, d\mathbb{P} + \int g \, d\mathbb{Q}.$$

Then, r is well-defined since if $f(x) + g(y) \equiv k(x) + j(y)$, then $f(x) - k(x) \equiv j(y) - g(y)$, which must be some constant c, so $k = f - c$ and $j = g + c$, giving $\int (f - k) \, d\mathbb{P} + \int (g - j) \, d\mathbb{Q} = 0$.

Kantorovich-Rubinstein theorem – proof

For any $h \in C(S \times S)$ let

$$U = U_h := \{k \in C(S \times S) : k(x, y) < h(x, y) \text{ for all } x, y\}.$$

Then, U is a convex set, open for the supremum norm (since S is compact).

Now r is a linear functional on L, not identically 0, and bounded above on the non-empty convex set $U \cap L$, since $f(x) + g(y) < h(x, y)$ for all x, y implies $r(\phi) \leq \sup(f) + \sup(g) < +\infty$.

So, by one form of the Hahn-Banach theorem, r can be extended to a linear functional ρ on $C(S \times S)$ with $\sup_{u \in U} \rho(u) = \sup_{v \in U \cap L} r(v)$.

Kantorovich-Rubinstein theorem – proof

Suppose $u \in C(S \times S)$ and $u(x,y) \geq 0$ for all x, y. Then, for any $c \geq 0$, $h - 1 - cu \in U$, so $\rho(h - 1 - cu)$ is bounded above as $c \to +\infty$, implying $\rho(u) \geq 0$.

By the Riesz representation theorem, there is a nonnegative, finite measure ρ on $S \times S$ such that

$$\rho(k) = \int k \, d\rho, \quad \text{for all } k \in C(S \times S).$$

Since $\rho = r$ on L, we have for any f and $g \in C(S)$ that

$$\int f(x) \, d\rho(x,y) = \int f \, d\mathbb{P} \quad \text{and} \quad \int g(y) \, d\rho(x,y) = \int g \, d\mathbb{Q}.$$

Thus, $\rho \in M(\mathbb{P}, \mathbb{Q})$.

Kantorovich-Rubinstein theorem – proof

Now,

$$m_{\mathbb{P},\mathbb{Q}}(h) = \sup_{u \in U \cap L} r(u) = \sup_{U} \rho(u) = \int h \, d\rho,$$

so

$$m_{\mathbb{P},\mathbb{Q}}(h) \geq \inf \left\{ \int h \, d\mu : \mu \in M(\mathbb{P}, \mathbb{Q}) \right\},$$

proving Lemma 2. $\square$

Kantorovich-Rubinstein theorem – proof

Lemma 3

If S is a compact metric space and $h \in C(S \times S)$ is a pseudometric on S, then

$$m_{\mathbb{P},\mathbb{Q}}(h) = \sup_{j \in J_h} \left| \int j \, d(\mathbb{P} - \mathbb{Q}) \right| = \sup_{j \in J_h} \int j \, d(\mathbb{P} - \mathbb{Q})$$

where

$$J_h := \{j \in C(S) : |j(x) - j(y)| \leq h(x,y) \text{ for all } x, y \in S\}.$$

Proof. If h is a pseudometric and $f(x) + g(y) < h(x,y)$ for all x and y, let

$$j(x) := \inf_{y \in S} (h(x,y) - g(y)).$$

Then, $f(x) < h(x,y) - g(y)$ for any $y \in S$, so $f \leq j$. Also, $j(x) \leq h(x,x) - g(x) = -g(x)$, so $j \leq -g$.

Kantorovich-Rubinstein theorem – proof

Moreover, for all x and x',

$$\begin{aligned} j(x) - j(x') &= \inf_{y \in S} (h(x,y) - g(y)) - \inf_{y' \in S} (h(x',y') - g(y')) \\ &= \sup_{y' \in S} \inf_{y \in S} (h(x,y) - g(y) + g(y') - h(x',y')) \\ &\leq \sup_{y' \in S} (h(x,y') - g(y') + g(y') - h(x',y')) \\ &= \sup_{y' \in S} (h(x,y') - h(x',y')) \leq h(x,x'), \end{aligned}$$

so $j \in J_h$ and $\int f \, d\mathbb{P} + \int g \, d\mathbb{Q} \leq \int j \, d(\mathbb{P} - \mathbb{Q})$.

Hence,

$$m_{\mathbb{P},\mathbb{Q}}(h) \leq \sup_{j \in J_h} \left| \int j \, d(\mathbb{P} - \mathbb{Q}) \right|.$$

The converse inequality always holds, since for any $j \in J_h$, we can let $f = -g = j$ in the definition of $m_{\mathbb{P},\mathbb{Q}}(h)$. $\square$

Kantorovich-Rubinstein theorem – proof

Completion of the proof of the Kantorovich-Rubinstein theorem (Theorem 1)

By Lemma 2 (with $h = d$), $W(\mathbb{P}, \mathbb{Q}) = m_{\mathbb{P},\mathbb{Q}}(d)$.

By Lemma 3 (with $h = d$), $m_{\mathbb{P},\mathbb{Q}}(d) = \gamma(\mathbb{P}, \mathbb{Q})$. $\square$

Lipschitz functions

It is instructive to prove directly that $\gamma(\mathbb{P}, \mathbb{Q}) = 0$ implies $\mathbb{P} = \mathbb{Q}$. This will follow if we can show that the set of Lipschitz functions is dense in $C(S)$ for the supremum norm $|| \cdot ||_\infty$.

The first step is the following.

Theorem 4 (Kirszbraun-McShane)

Let (S, d) be a metric space, E any subset of S, and f any Lipschitz function on E. Then, f can be extended to all of S without increasing $||f||_L$.

Lipschitz functions

Proof. Set $M := ||f||_L$. Suppose we have an inclusion-chain of functions f_α from subsets E_α of S into $\mathbb{R}$ with $||f_\alpha||_L \leq M$ for all α. Let g be the union of the f_α. Then, g is a function with $||g||_L \leq M$.

Thus, by Zorn's Lemma it suffices to extend f to one additional point $x \in S \setminus E$.

A real number y is a possible value of $f(x)$ if and only if $|y - f(u)| \leq Md(u, x)$ for all $u \in E$, or equivalently the following two conditions both hold:

(i) $-Md(u, x) \leq y - f(u)$ for all $u \in E$, and

(ii) $y - f(v) \leq Md(x, v)$ for all $v \in E$.

Such a y exists if and only if

$$\sup_{u \in E}(f(u) - Md(u,x)) \leq \inf_{v \in E}(f(v) + Md(v,x)). \tag{1}$$

Now, by assumption, for all $u, v \in E$,

$$f(u) - f(v) \leq Md(u,v) \leq Md(u,x) + Md(x,v),$$

$$f(u) - Md(u,x) \leq f(v) + Md(x,v).$$

This implies (1), so the extension to x is possible. $\square$

Lipschitz functions

Theorem 5

If (S, d) is a compact metric space, then the set of Lipschitz functions is dense in $C(S)$ for the supremum norm $||\cdot||_\infty$.

Proof. The set of Lipschitz functions is clearly an algebra. It separates points by Theorem 4 (taking E to be a two-point set and f any non-constant function on E). Thus, the Stone-Weierstrass theorem applies. □

"Zolotarev" distances

The formula

$$Z_1(\mathbb{P},\mathbb{Q}) = \int_{\mathcal{T}} |\mathbb{P}(\mathcal{S}(y)) - \mathbb{Q}(\mathcal{S}(y))| \ \lambda(dy)$$

suggests defining a family of metrics by

$$Z_p(\mathbb{P},\mathbb{Q}) := \left\{ \int_{\mathcal{T}} |\mathbb{P}(\mathcal{S}(y)) - \mathbb{Q}(\mathcal{S}(y))|^p \ \lambda(dy) \right\}^{\frac{1}{p}}, \quad p \geq 1,$$

and

$$Z_p(\mathbb{P},\mathbb{Q}) := \left\{ \int_{\mathcal{T}} |\mathbb{P}(\mathcal{S}(y)) - \mathbb{Q}(\mathcal{S}(y))|^p \ \lambda(dy) \right\}, \quad 0 < p < 1.$$

Does the choice of ρ matter?

If ρ' and ρ'' are two choices of the "root" and $[\rho', \rho'']$ is the path in $\mathcal{T}$ joining them, then (in an obvious notation)

$$\mathbb{P}(\mathcal{S}'(y)) - \mathbb{Q}(\mathcal{S}'(y)) = \mathbb{P}(\mathcal{S}''(y)) - \mathbb{Q}(\mathcal{S}''(y)), \quad y \notin [\rho', \rho''],$$

and

$$\mathbb{P}(\mathcal{S}'(y)) - \mathbb{Q}(\mathcal{S}'(y)) = -\left[\mathbb{P}(\mathcal{S}''(y)) - \mathbb{Q}(\mathcal{S}''(y))\right], \quad \lambda - \text{a.e. } y \in [\rho', \rho''];$$

so

$$\int_{\mathcal{T}} \left|\mathbb{P}(\mathcal{S}'(y)) - \mathbb{Q}(\mathcal{S}'(y))\right|^p \, \lambda(dy) = \int_{\mathcal{T}} \left|\mathbb{P}(\mathcal{S}''(y)) - \mathbb{Q}(\mathcal{S}''(y))\right|^p \, \lambda(dy)$$

and the choice of the "root" is irrelevant.

Assessing significance – a "permutation test"

- Recall $\mathbb{P} := \frac{1}{m}\sum_i \mathbb{P}_i$ and $\mathbb{Q} = \frac{1}{n}\sum_j \mathbb{Q}_j$.
- Is the observed value of $Z_p(\mathbb{P}, \mathbb{Q})$ "significant"?
- - Set $\{\mathbb{R}_1, \ldots, \mathbb{R}_{m+n}\} = \{\mathbb{P}_1, \ldots, \mathbb{P}_m\} \cup \{\mathbb{Q}_1, \ldots, \mathbb{Q}_n\}$.
  - Let I be a uniform random subset of $\{1, \ldots, m+n\}$ and put $J := \{1, \ldots, m+n\} \setminus I$.
  - Define

$$\tilde{\mathbb{P}} := \frac{1}{m}\sum_{i \in I} \mathbb{R}_i$$

and

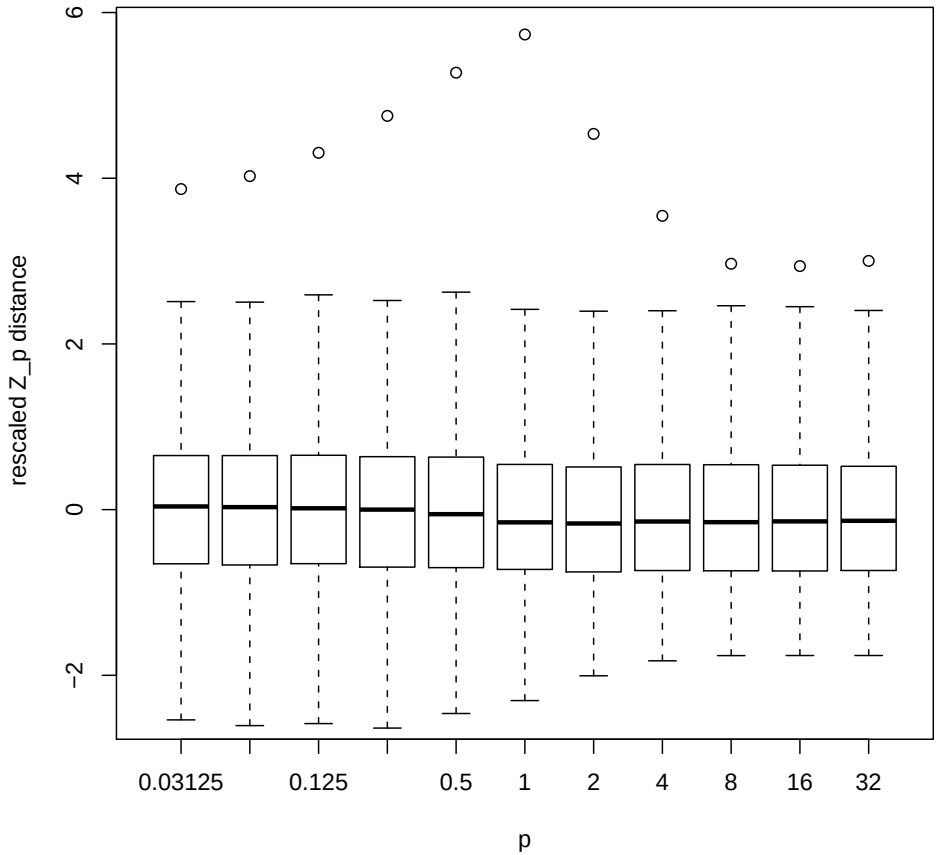
$$\tilde{\mathbb{Q}} := \frac{1}{n}\sum_{j \in J} \mathbb{R}_j.$$

  - Consider

$$\mathbf{Prob}\{Z_p(\tilde{\mathbb{P}}, \tilde{\mathbb{Q}}) > Z_p(\mathbb{P}, \mathbb{Q})\}.$$

- We can approximate this probability using Monte Carlo.

Choice of p?



Box plots of distributions of $Z_p(\tilde{\mathbb{P}}, \tilde{\mathbb{Q}})$ for the sea-water data set standardized to have mean 0 and variance 1 and observed $Z_p(\mathbb{P}, \mathbb{Q})$ (circles).

Gaussian approximation

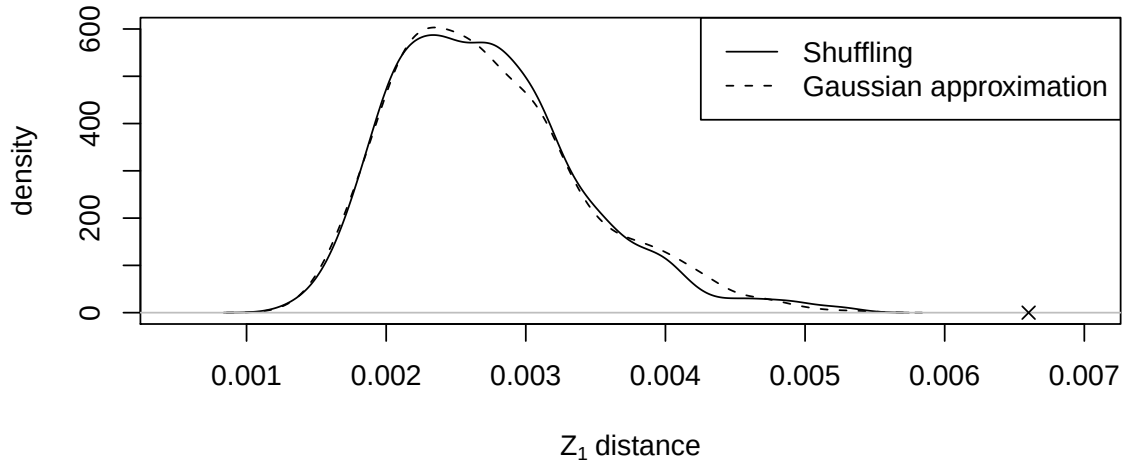
- Recall $Z_p(\tilde{\mathbb{P}}, \tilde{\mathbb{Q}}) = \left\{ \int_{\mathcal{T}} |X(u)|^p \, \lambda(du) \right\}^{\frac{1}{p}}$, where $X(u) := \frac{1}{m} \sum_{i \in I} G_i(u) - \frac{1}{n} \sum_{j \in J} G_j(u)$ with $G_k(u) := \mathbb{R}_k(\mathcal{S}(u))$. Put $\bar{G}(u) := \frac{1}{m+n} \sum_k G_k(u)$.
- When m, n are large, $(X(u))_{u \in \mathcal{T}}$ is approximately a mean zero Gaussian process with covariance kernel

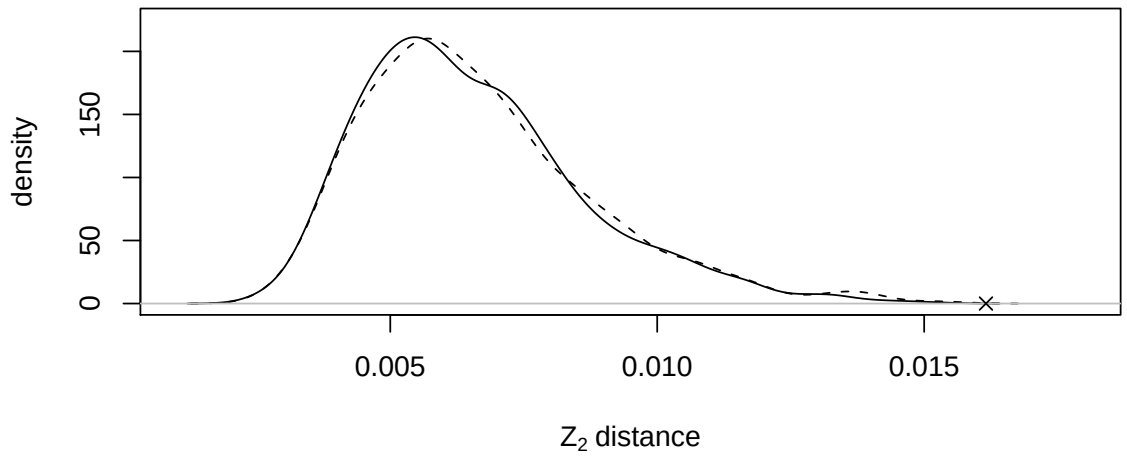
$$\Gamma(u, v) := \frac{1}{mn} \sum_k (G_k(u) - \bar{G}(u))(G_k(v) - \bar{G}(v)).$$

- We may construct a Gaussian process $(\xi(u))_{u \in \mathcal{T}}$ with covariance kernel Γ by taking independent standard normal random variables $\zeta_1, \ldots, \zeta_{m+n}$ and setting

$$\xi(u) := \frac{1}{\sqrt{mn}} \left[\sum_i G_i(u) \zeta_i - \frac{1}{m+n} \left(\sum_i G_i(u) \right) \left(\sum_i \zeta_i \right) \right].$$

How good is the approximation? $(p = 1, 2)$





Understanding Z_2

- Recall that $Z_2^2(\tilde{\mathbb{P}}, \tilde{\mathbb{Q}}) \approx \int_{\mathcal{T}} \xi(u)^2 \, \lambda(du)$, where $(\xi(u))_{u \in \mathcal{T}}$ is Gaussian with covariance kernel

$$\Gamma(u, v) := \frac{1}{mn} \sum_k (G_k(u) - \bar{G}(u))(G_k(v) - \bar{G}(v)).$$

- Thus, $Z_2^2(\tilde{\mathbb{P}}, \tilde{\mathbb{Q}}) \approx \sum_k \mu_k^2 \eta_k^2$, where μ_k^2 are the non-zero eigenvalues of the linear operator

$$\Gamma f(x) := \int_{\mathcal{T}} \Gamma(x, y) f(y) \, \lambda(dy)$$

on $L^2(\lambda)$ and η_k are i.i.d. standard normal (Karhunen-Loève).

Simplifying Z_2

- The non-zero eigenvalues of Γ are the same as those of the $(m+n) \times (m+n)$ matrix

$$M_{ij} := \frac{1}{mn} \int_{\mathcal{T}} \left(G_i(u) - \bar{G}(u)\right) \left(G_j(u) - \bar{G}(u)\right) \lambda(du).$$

- Thus,

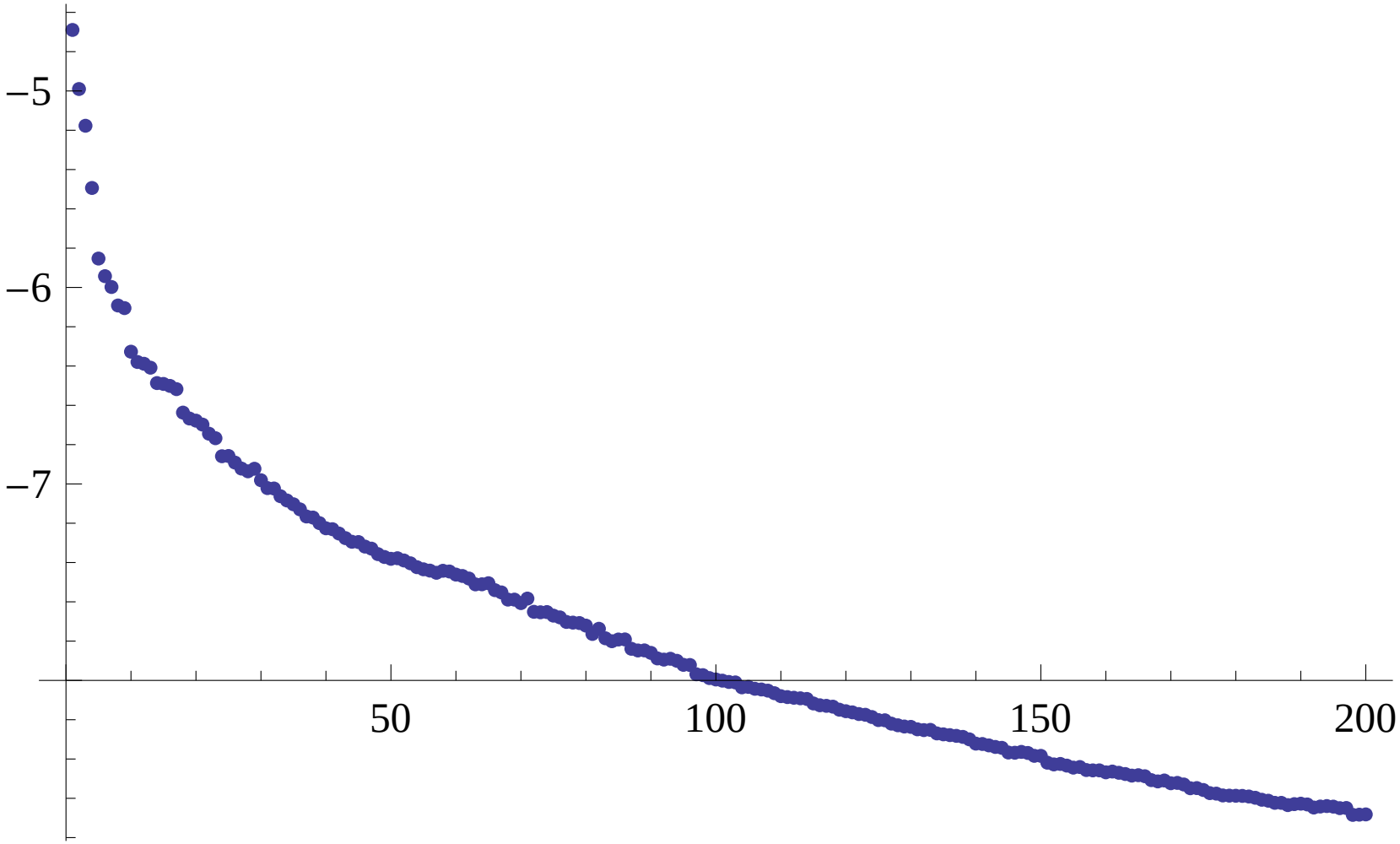
$$Z_2^2(\tilde{\mathbb{P}}, \tilde{\mathbb{Q}}) \approx \int_{\mathcal{T}} \xi(u)^2 \, \lambda(du) = \sum_k \mu_k^2 \eta_k^2$$

has the same distribution as

$$\eta^T M \eta,$$

where η is a vector of i.i.d. standard normals.

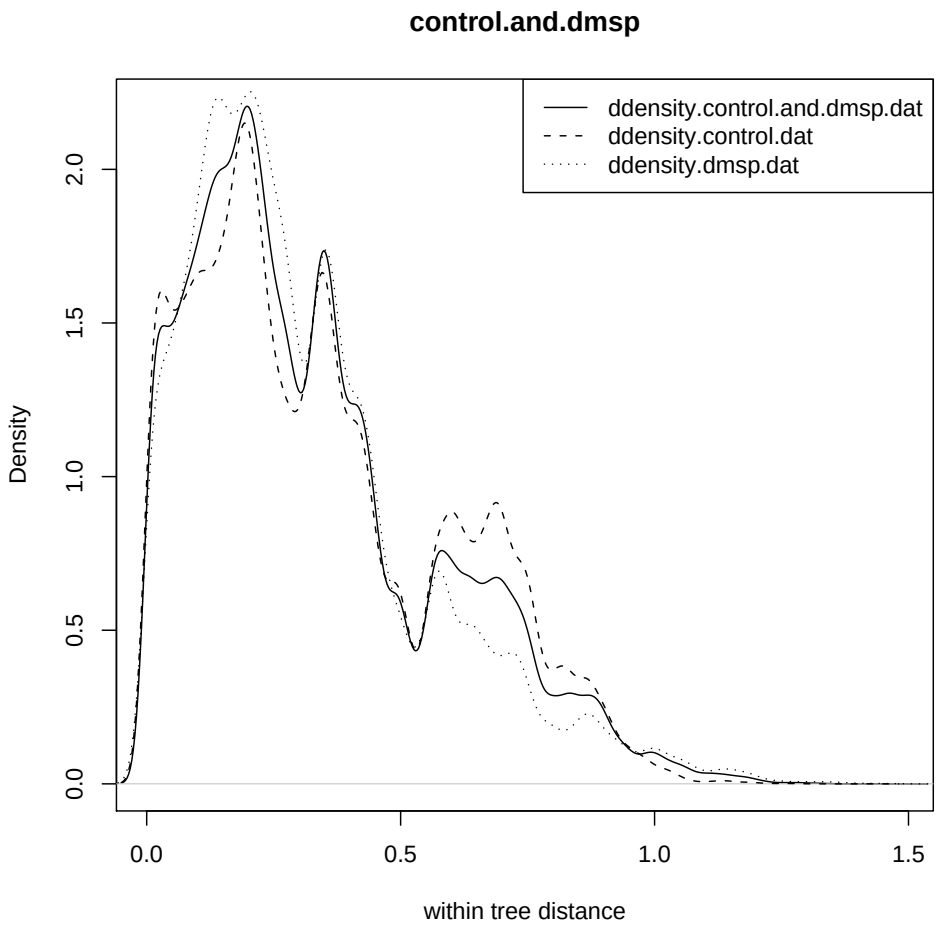
Eigenvalues of Γ and M



Log base 10 of the first 200 eigenvalues of Γ and M.

Other ways to look at these data?

Investigate a probability measure $\mathbb{P}$ on $\mathcal{T}$ by looking at the push-forward of $\mathbb{P} \otimes \mathbb{P}$ by the map $(u, v) \mapsto d(u, v)$.



Metagenomics review

- Next-generation sequencing technology enables sequencing of hundreds of thousands to millions of short DNA sequences in a single experiment.
- Microbial genetic material can be extracted in bulk from a sample taken from some environment and directly sequenced.
- It is no longer necessary to identify individual species by morphology or culturing experiments.
- This technology has revolutionized the possibilities for surveys of environmental microbial diversity.

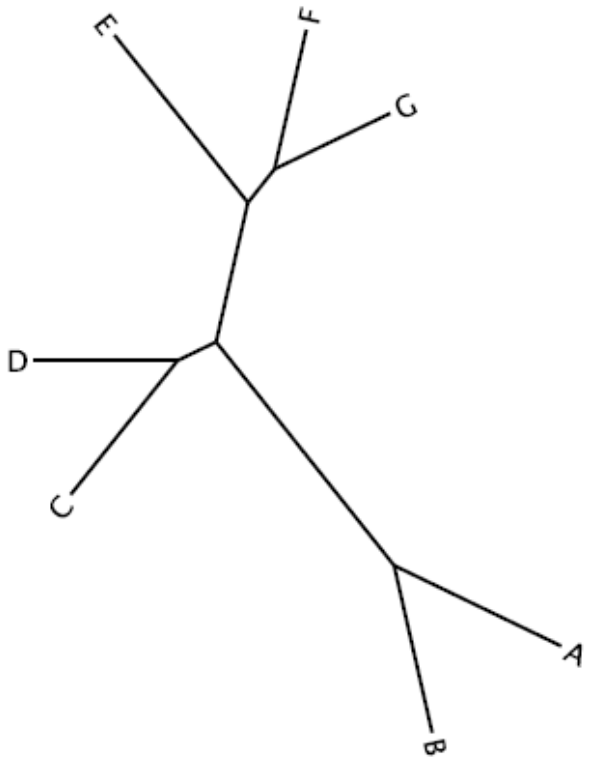
Metagenomics review – continued

- Huge amount of data – 400K sequences from a single run.
- "Reads" (little shreds of DNA) are often short and non-overlapping.
- Not enough signal in the data to resolve a phylogenetic tree.
- Computationally infeasible to build a tree anyway.

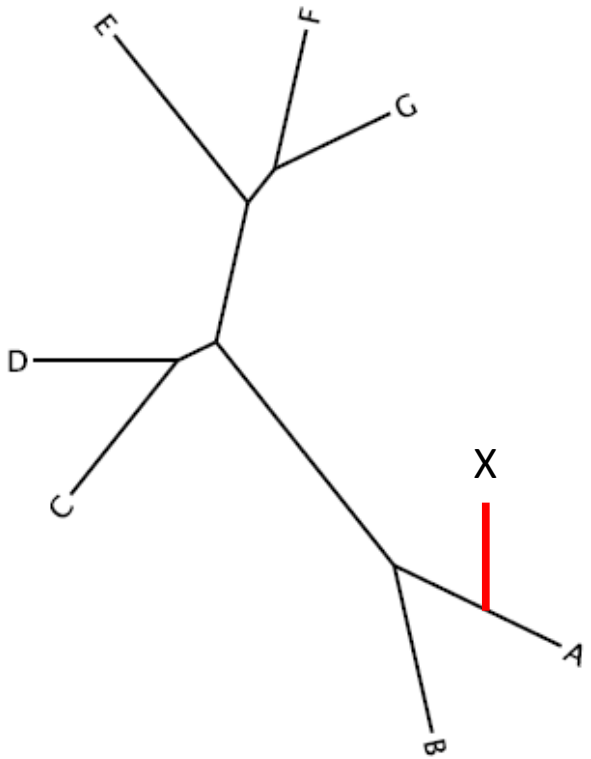
Metagenomics review – continued

- Begin with a reference phylogenetic tree constructed from previously-characterized DNA sequences.
- Place query sequences at their most likely position on the reference tree.
- Recent such algorithms are able to place tens of thousands of query sequences on a reference tree of one thousand taxa (species).
- Rather than just point placements we can also have individual placements that are spread out as a probability distribution on the reference tree.

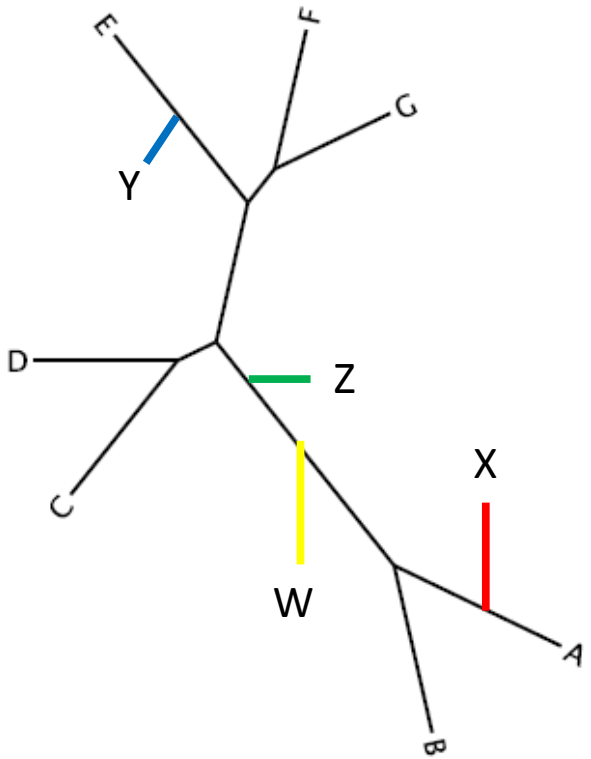
Metagenomics review – continued



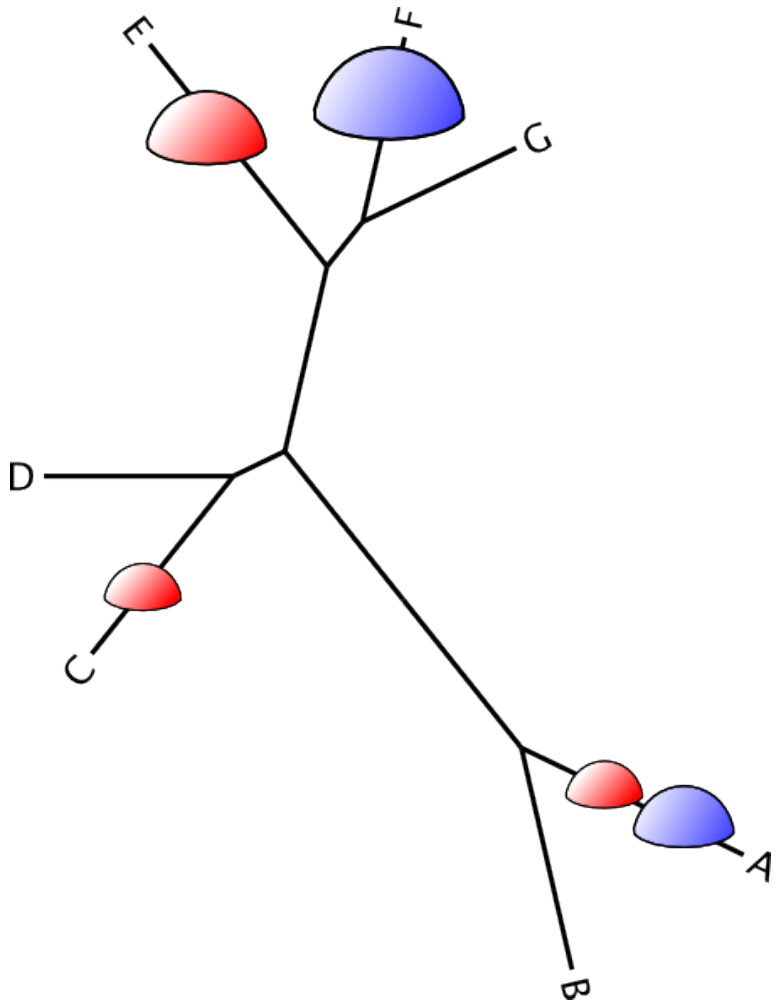
Metagenomics review – continued



Metagenomics review – continued



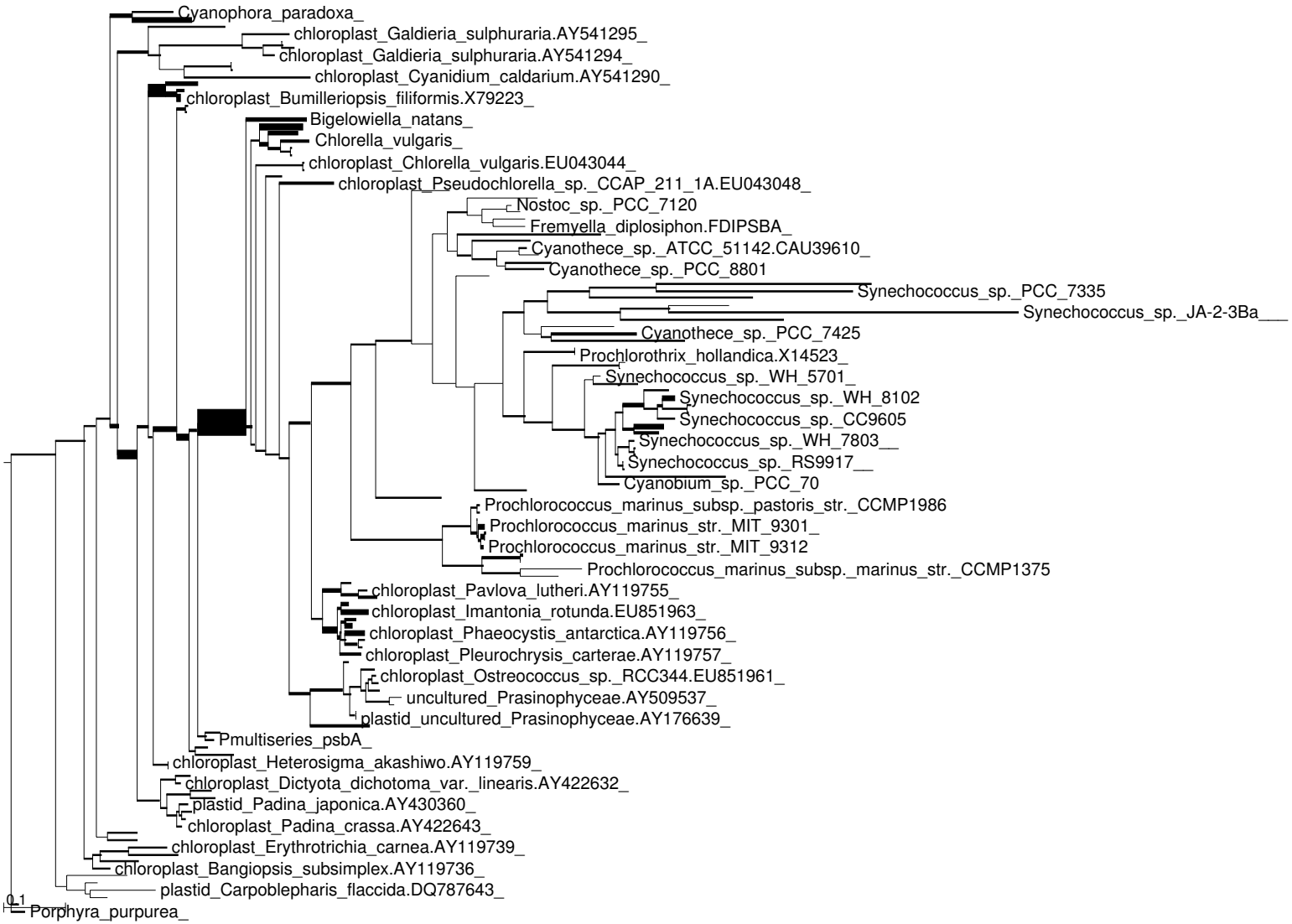
Metagenomics review – continued



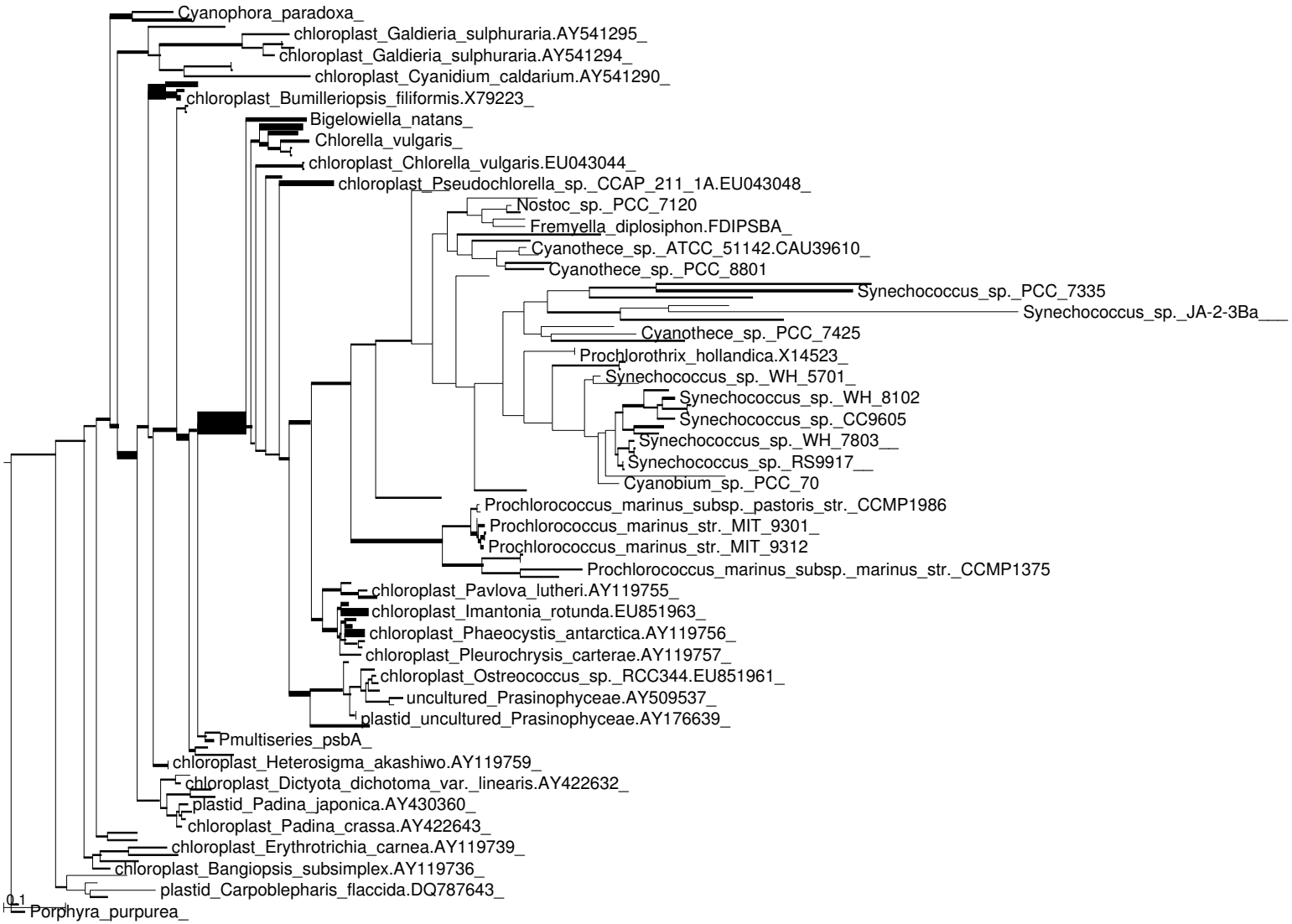
Metagenomics review – continued

- We have m placements in a sample.
- The i^{th} placement corresponds to a probability measure $\mathbb{P}_i$ on the reference tree $\mathcal{T}$.
- We represent the entire sample using the probability measure $\mathbb{P} := \frac{1}{m} \sum_i \mathbb{P}_i$.

Metagenomics review – continued



Metagenomics review – continued



Several metagenomic samples?

HOW DO WE VISUALIZE AND COMPARE MORE THAN TWO METAGENOMIC SAMPLES?

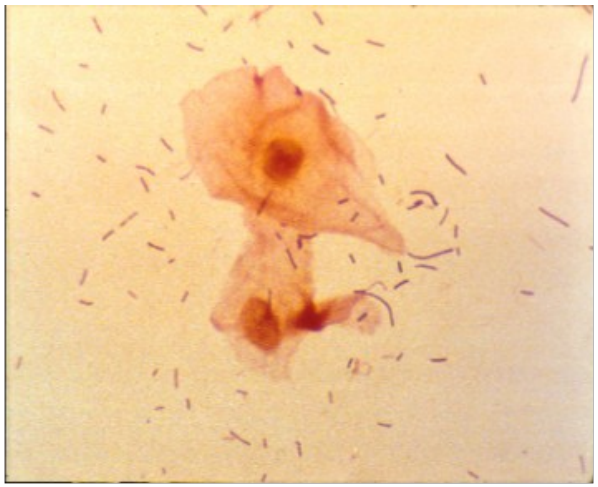
An example

- Swabs were taken from 242 women from the Public Health, Seattle and King County Sexually Transmitted Diseases Clinic (STD clinic) between September 2006 and June 2010 (of which 222 samples resulted in enough material to analyze).
- DNA was extracted and the 16s gene was amplified in the V3-V4 hypervariable region using broad-range primers and sequenced using a 454 sequencer with FLX chemistry.
- Sequences were pre-processed using the R / Bioconductor package *microbiome*. A custom maximum likelihood reference tree consisting of sequences from RDP and our local collection was built using *RAxML* 7.2.7 using GTR+4Γ.
- Sequences were placed into this tree using *pplacer* with the default parameter choices.

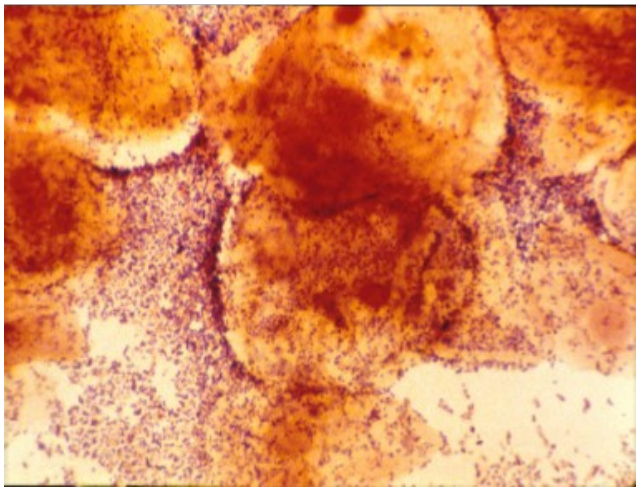
What Wikipedia says about bacterial vaginosis

- Bacterial vaginosis (BV) is the most common cause of vaginal infection.
- It is not considered to be a sexually transmitted infection by the CDC.
- BV is not transmitted through sexual intercourse but is more common in women who are sexually active.
- BV is caused by an imbalance of naturally occurring bacterial flora and is often confused with yeast infection.
- The diagnostic standard for researchers is the Nugent test. A score of 0-10 is generated from combining three other scores that reflect the prevalence of bacteria with certain shapes and reaction to staining.
  - 0-3 is considered negative for BV
  - 4-6 is considered intermediate
  - 7+ is considered indicative of BV

Bacterial Vaginosis (BV)



Gram stain of normal vaginal fluid with many GPR (lactobacilli), normal epithelial cells

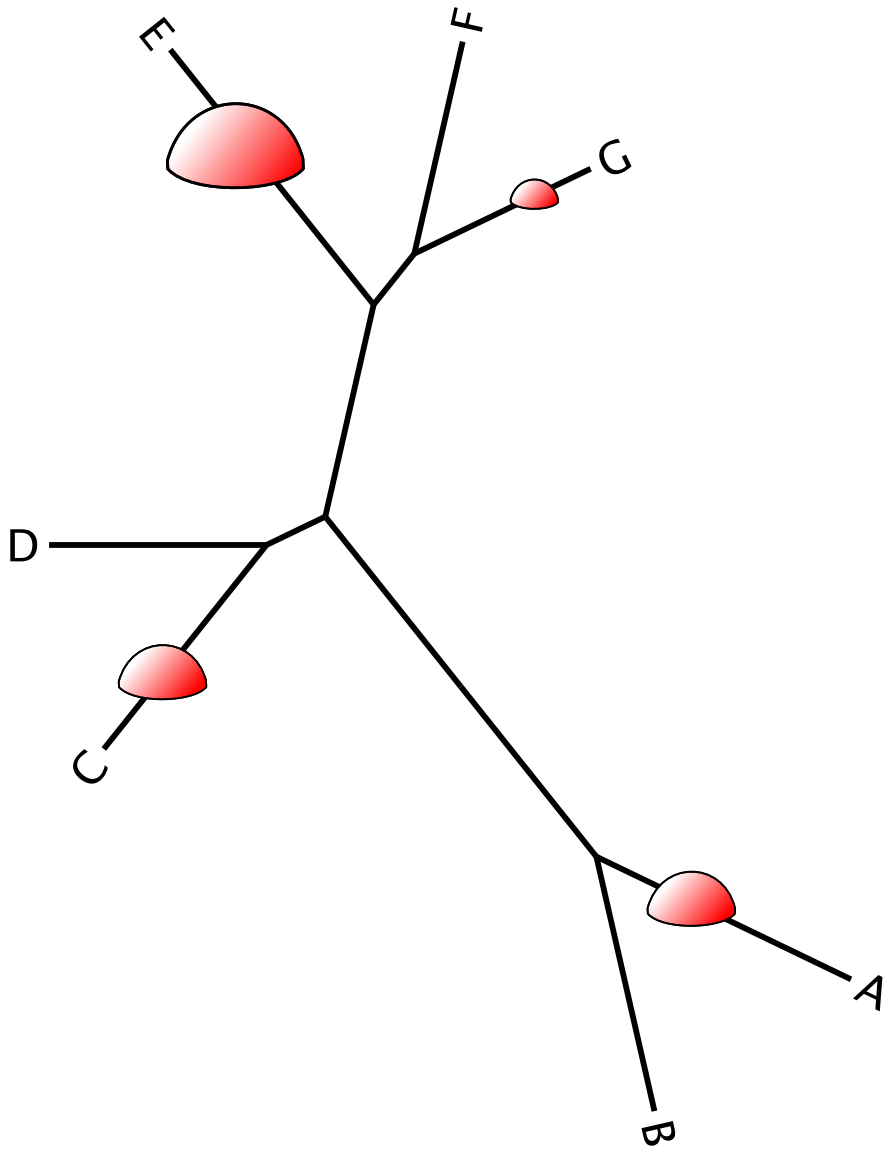


Gram stain of BV with few GPR, greater diversity of morphotypes, and clue cells

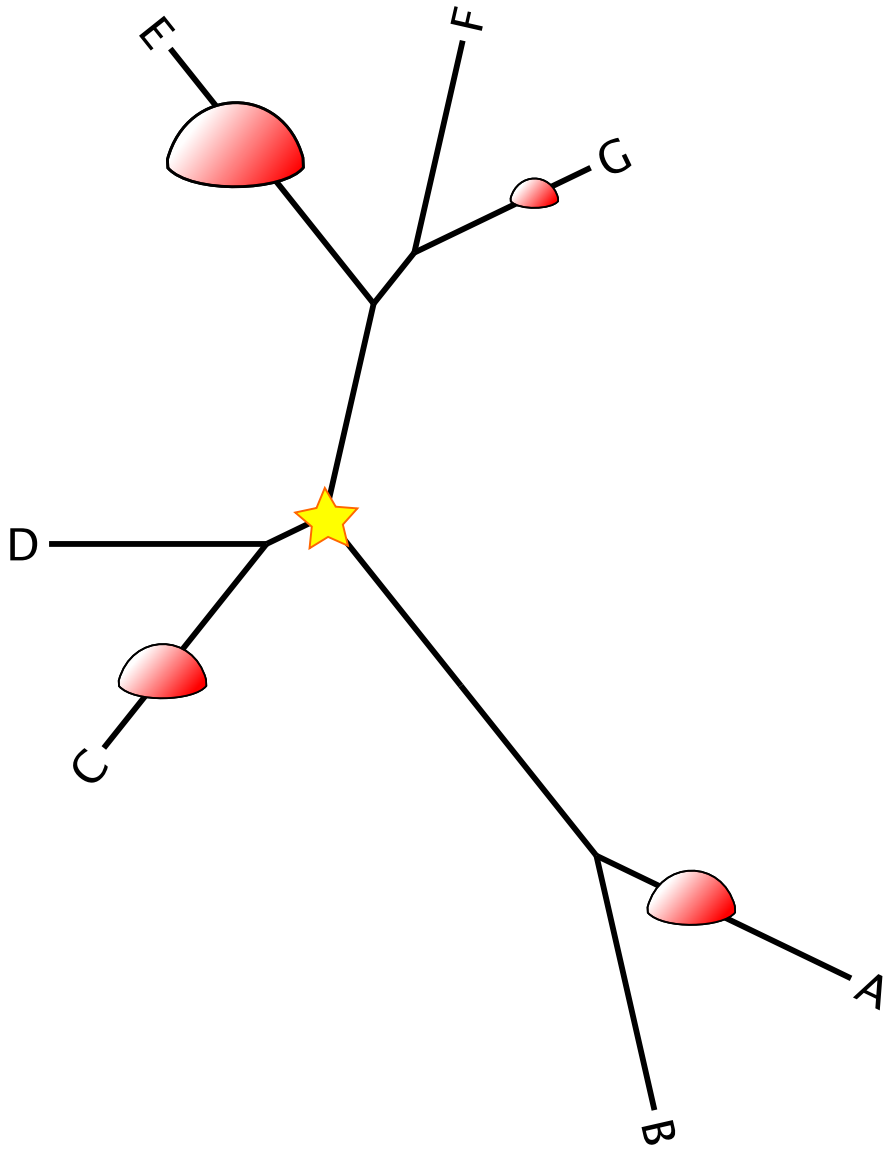
Metagenomic data as vectors indexed by internal edges

- Suppose we have S metagenomic samples.
- Each sample is encoded as a mass distribution on a reference tree with E internal edges.
- Distinguish an arbitrary vertex of the tree as the root and map each mass distribution to an E-dimensional vector by recording for each internal edge the difference between the total mass on the root side of the edge and the total mass on the non-root side of the edge.
- This results in an $S \times E$ "data matrix".

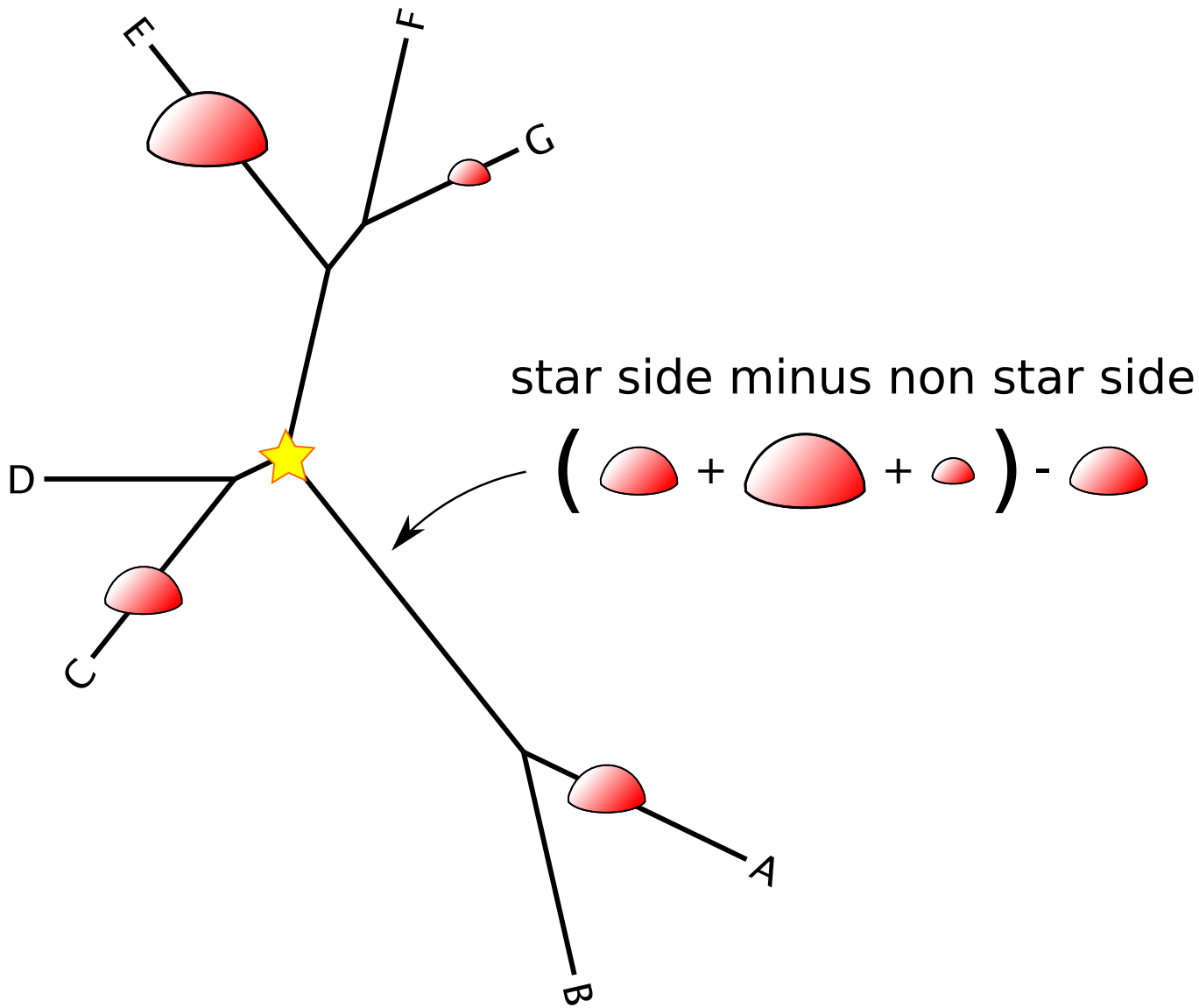
Metagenomic data as vectors indexed by internal edges – continued



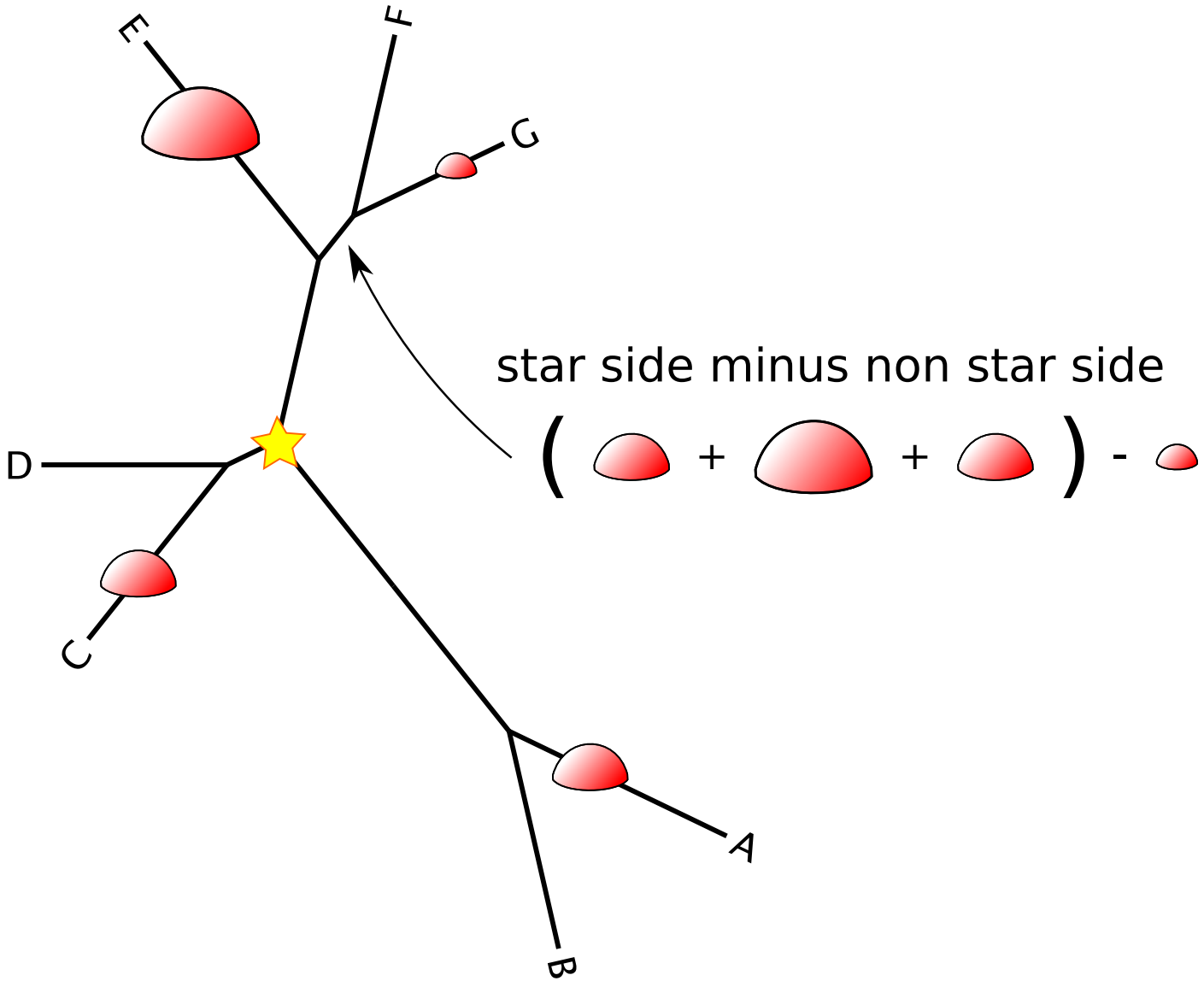
Metagenomic data as vectors indexed by internal edges – continued



Metagenomic data as vectors indexed by internal edges – continued



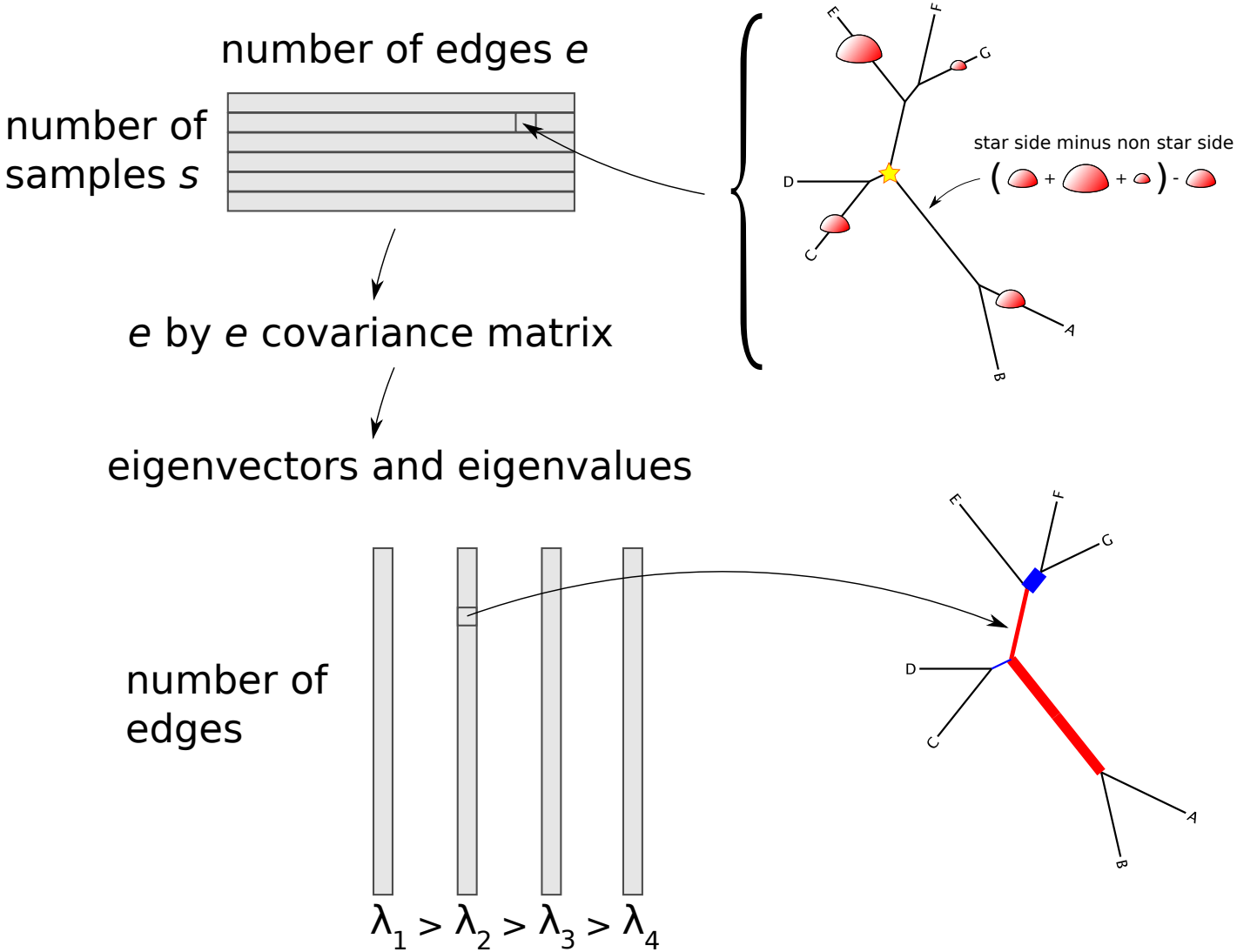
Metagenomic data as vectors indexed by internal edges – continued



Edge principal components

- Recall we had S metagenomic samples.
- Each sample was encoded as a mass distribution on a reference tree with E internal edges.
- We transformed each mass distribution into a vector indexed by the internal edges to get an $S \times E$ "data matrix".
- Now construct the $E \times E$ covariance matrix of this data matrix.
- Calculate its eigenvalues and their corresponding eigenvectors.
- An eigenvector can be represented by a single colored and thickened reference tree: the thickness of an edge is proportional to the magnitude of the corresponding entry of the eigenvector and the color specifies the sign of that entry.

Edge principal components schematic



What do eigenvalues and eigenvectors mean?

- Recall the variational characterization of the eigenvectors $v_1, \ldots, v_E$ of an $E \times E$ non-negative definite matrix Σ listed in order of decreasing eigenvalue:

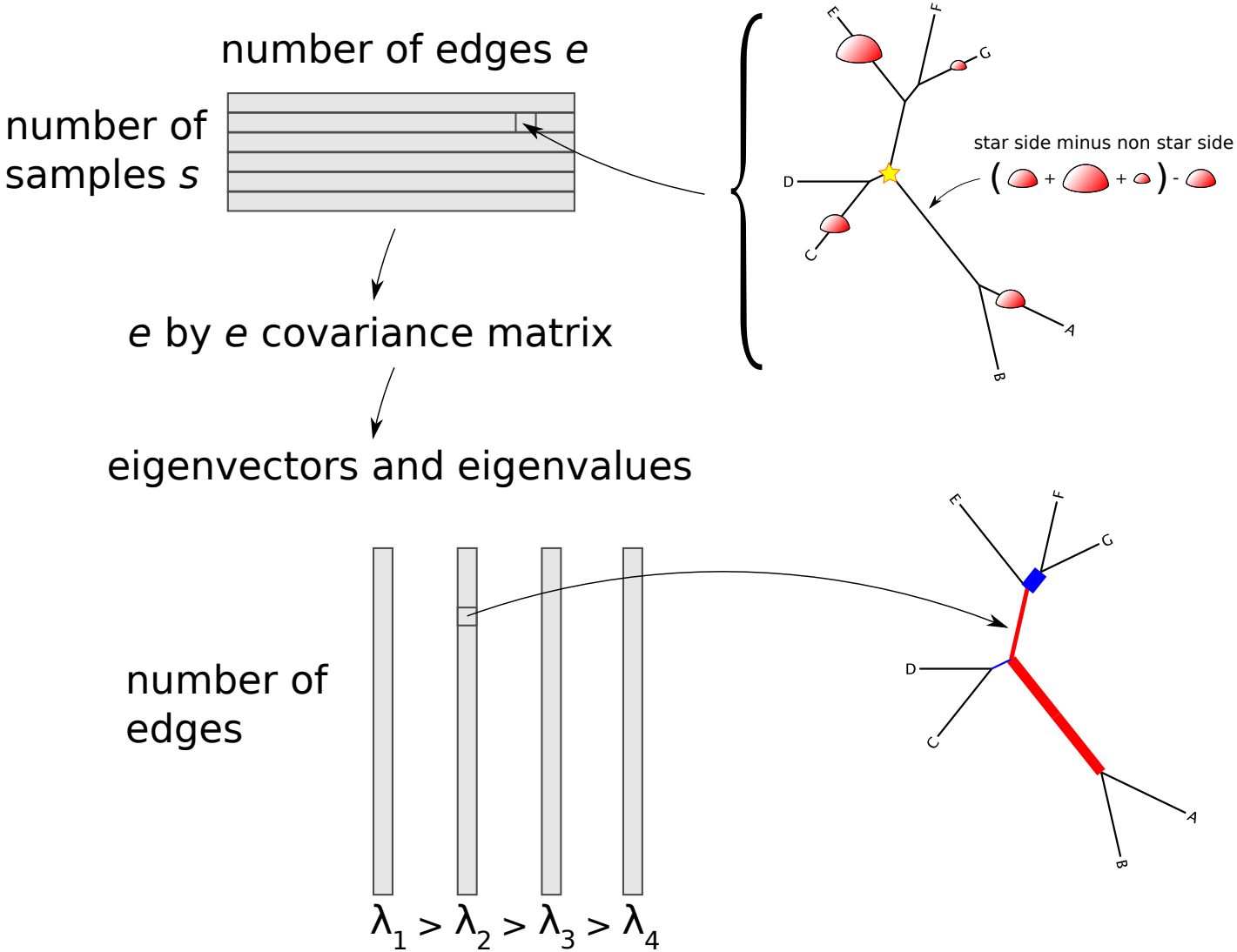
$$
\begin{aligned}
v_1 &= \arg\max_{||v||=1} \langle v, \Sigma v \rangle \\
v_2 &= \arg\max_{||v||=1, v \perp v_1} \langle v, \Sigma v \rangle \\
&\ldots \\
v_E &= \arg\max_{||v||=1, v \perp \{v_1, \ldots, v_{E-1}\}} \langle v, \Sigma v \rangle,
\end{aligned}
$$

where $\|v\|$ is the Euclidean length of the vector v, $\langle v, w \rangle$ is the Euclidean inner product of the vectors v and w, and $v \perp \{v_1, \ldots, v_k\}$ indicates that v is perpendicular to each of the vectors $v_1, \ldots, v_k$.

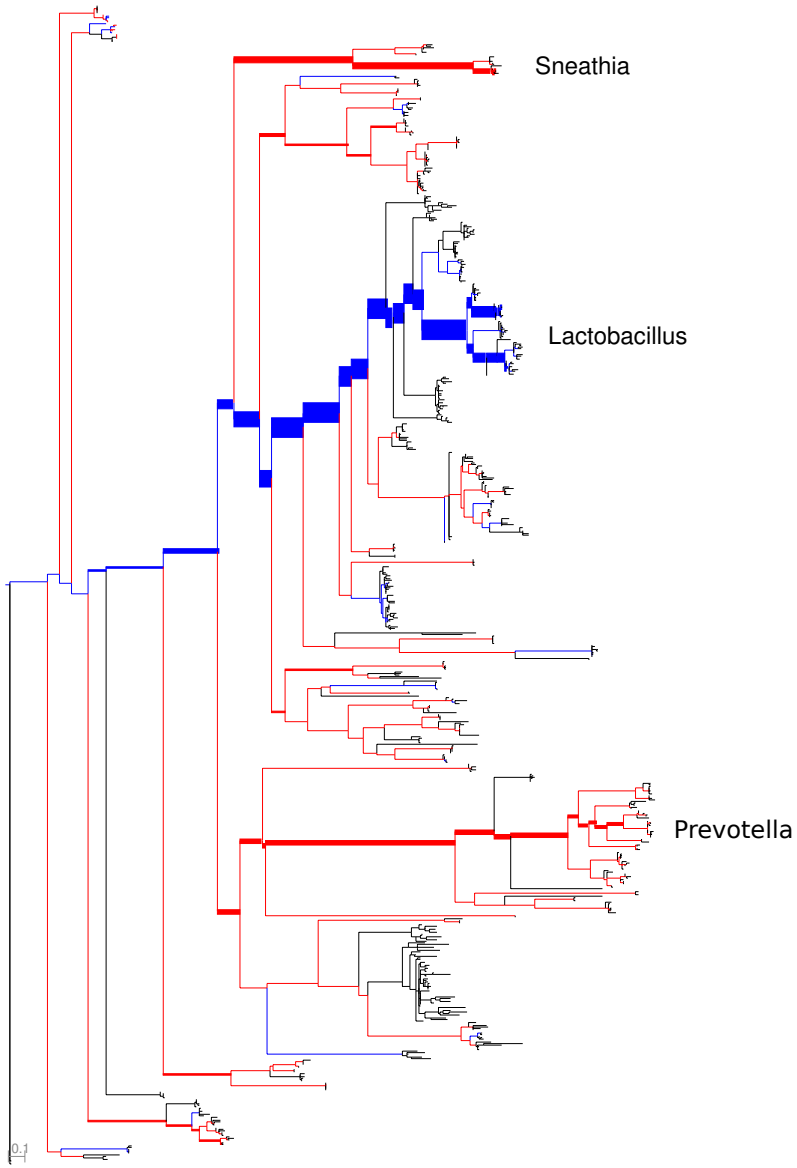
Interpretation of edge principal components

- Thus, an edge that receives a weight with large magnitude from an eigenvector corresponding to one of the bigger eigenvalues of the covariance matrix Σ may be viewed as an edge for which there are substantial dissimilarities between samples in the amount of mass placed in the components of the tree $\mathcal{T}$ on either side of the edge.
- When looking at the weight assigned to a single edge in isolation, only the magnitude matters and not the sign.
- Changing the chosen root from ρ' to vertex ρ'' does not affect the eigenvalues or the magnitudes of the entries in the corresponding eigenvectors, and it only changes the signs of the eigenvalue entries for the edges between ρ' and ρ''.

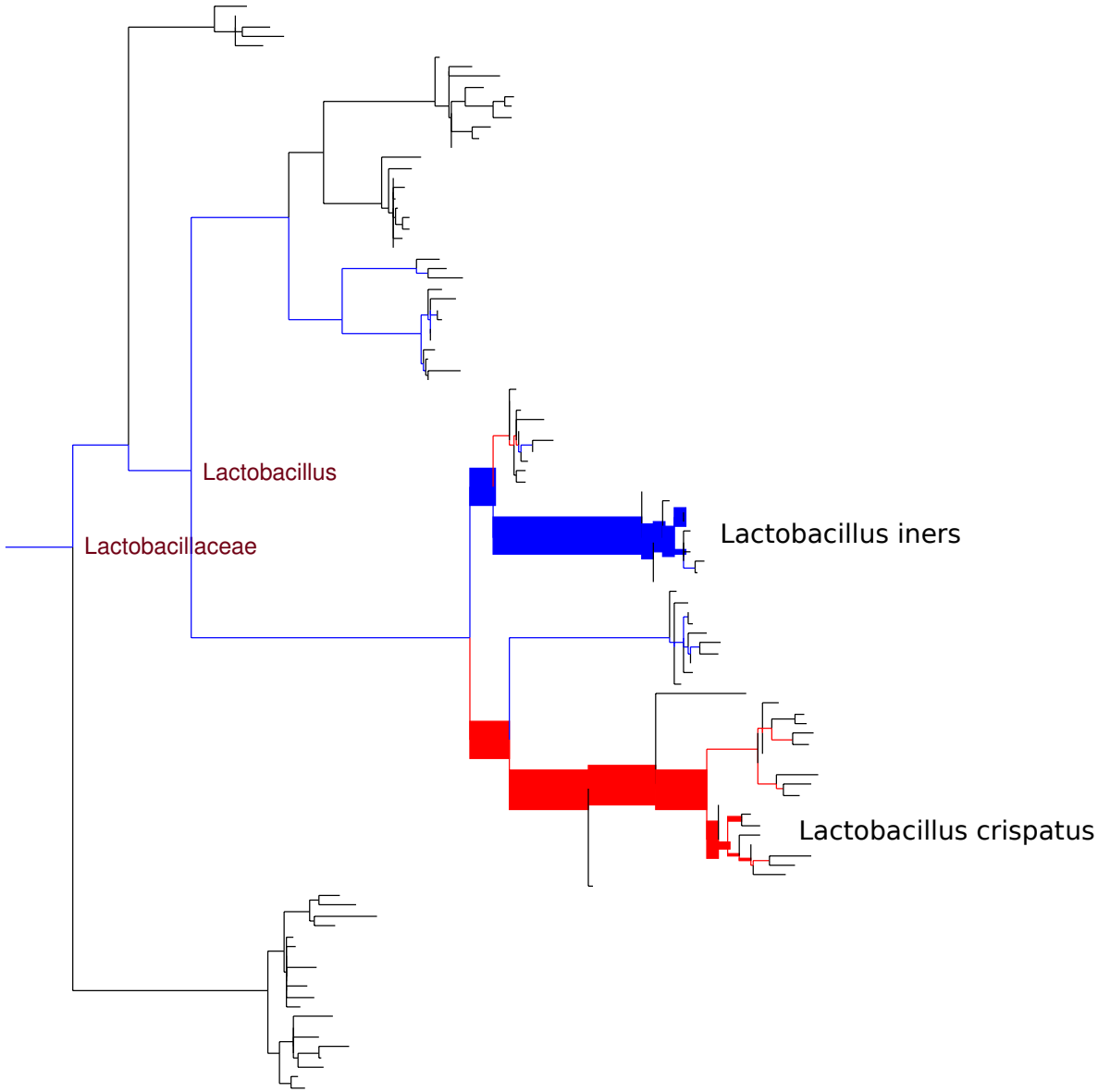
Edge principal components – reminder



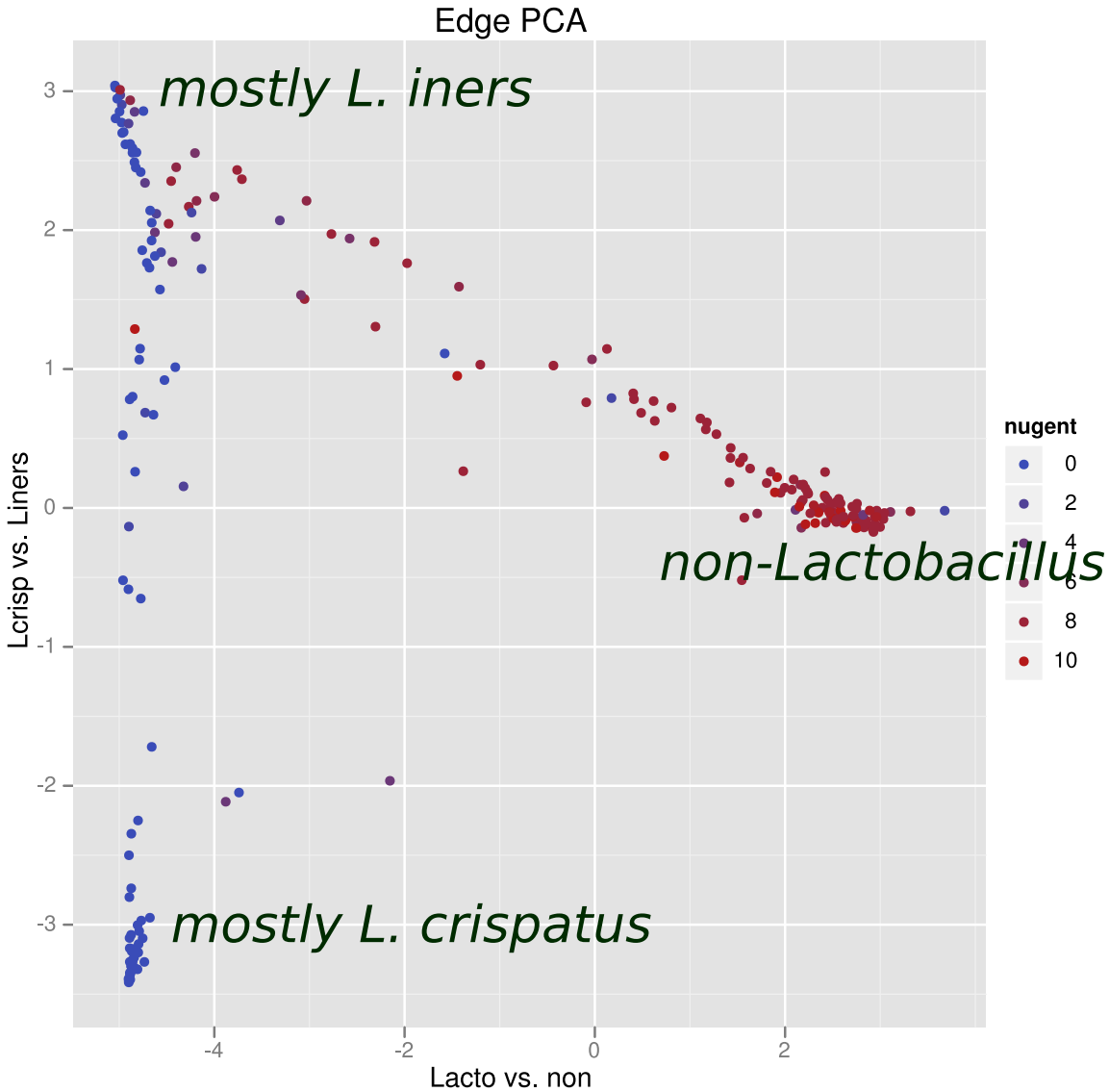
First edge principal component for example



Second edge principal component for example



What do the first two edge principal components tell us?



Detecting hierarchical structure

- In the BV example, we had at least 3 clusters of samples:
  - *Lactobacillus iners* predominant
  - *Lactobacillus crispatus* predominant
  - Others (e.g. *Sneathia* and *Prevotella*) predominant
- Moreover, the first two clusters formed a "supercluster".
- HOW CAN WE DETECT THIS SORT OF HIERARCHICAL CLUSTERING?

Squash clustering

- At each stage of the following clustering algorithm we have a pairwise distance matrix with rows and columns indexed by the clusters that have already been made by the algorithm.
- Initially, the clusters are just the individual samples and the entries in the pairwise distance matrix are computed using the $Z_p(\cdot, \cdot)$ distance.
- The algorithm proceeds by iterating the following sequence of steps until there is a single cluster and a corresponding 1×1 pairwise distance matrix.
  1. Find the smallest off-diagonal element in the current pairwise distance matrix. Say it is the distance between clusters i and j.
  2. Merge the i and j clusters, making a new cluster k.
  3. Remove the ith and jth rows and columns from the distance matrix.
  4. Calculate the distance from the cluster k to all the other clusters (how???).
  5. Insert the distances from k into the distance matrix.

How do we update the distances?

- At each stage of squash clustering, a cluster is associated with a probability measure on the reference tree $\mathcal{T}$.
- When two clusters i and j containing respective numbers of items a and b and associated with respective probability measures $\mathbb{P}$ and $\mathbb{Q}$ are merged to form a cluster k, then the new cluster k is associated with the probability measure $\frac{a}{a+b}\mathbb{P} + \frac{b}{a+b}\mathbb{Q}$ and the distance from k to some other cluster ℓ associated with the probability measure $\mathbb{R}$ is

$$Z_p\left(\frac{a}{a+b}\mathbb{P} + \frac{b}{a+b}\mathbb{Q}, \mathbb{R}\right)$$

Updating distances – continued

- That is, if S_z is the probability measure associated with original item z, then

$$\mathbb{P} = \frac{1}{a} \sum_{x \in i} S_x$$

and

$$\mathbb{Q} = \frac{1}{b} \sum_{y \in j} S_y,$$

and the probability measure associated with the new cluster k is

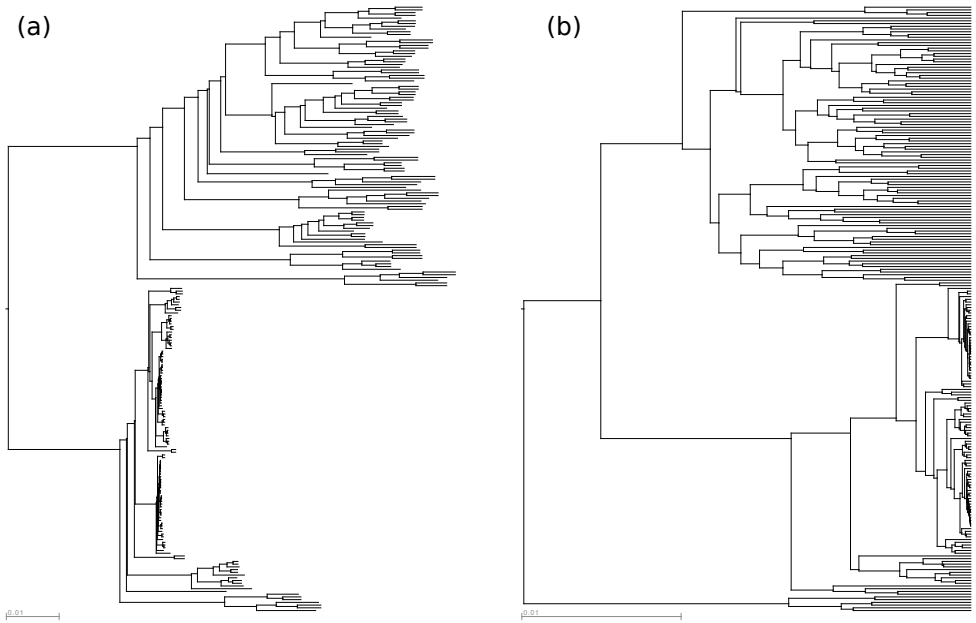
$$\frac{a}{a+b}\mathbb{P} + \frac{b}{a+b}\mathbb{Q} = \frac{1}{a+b} \sum_{z \in k} S_z.$$

UPGMA

- Average-linkage clustering or UPGMA unweighted pair group method with arithmetic mean is a classical general purpose clutering algorithm that looks similar to squash clustering.
- UPGMA calculates the distance between two clusters as the average between pairs of items in the clusters.
- Thus, if clusters i and j containing respective numbers of items a and b are merged to form a new cluster k with $a+b$ items, then the distance between another cluster ℓ with c items and the new cluster k is

$$\begin{aligned}\text{distance}(k,\ell) &= \frac{1}{(a+b)c}\sum_{y\in k, z\in \ell} d(y,z)\\ &= \frac{a}{a+b}\frac{1}{ac}\sum_{w\in i, z\in \ell} d(w,z) + \frac{b}{a+b}\frac{1}{bc}\sum_{x\in j, z\in \ell} d(x,z)\\ &= \frac{a}{a+b}\text{distance}(i,\ell) + \frac{b}{a+b}\text{distance}(j,\ell).\end{aligned}$$

Squash clustering vs UPGMA



Squash clustering and UPGMA applied to the collection of vaginal samples
Because meaningful internal edge lengths can be assigned to the squash clustering tree, it is not ultrametric, whereas the UPGMA tree is. The two tight clusters at the bottom of (a) and (b) contain the *Lactobacillus*-dominated vaginal samples and correspond to *L. iners* (upper tight cluster) and *L. crispatus* (lower tight cluster).