Phylogeny-based methods for analysing genomes and metagenomes

Presented by Aaron Darling
OTUs for ecology

Operational Taxonomic Unit: a grouping of similar sequences that can be treated as a single “species”

- **Strengths**
  - Conceptually simple
  - Mask effect of poor quality data
    - Sequencing error
    - *in vitro* recombination

- **Weaknesses**
  - Limited resolution
  - Logically inconsistent definition
Logical inconsistency: OTUs at 97% ID

Assume the true phylogeny:

OTU pipelines will arbitrarily pick one of the three solutions. Is this actually a problem??

Possible valid OTUs:
AB, C (with A & C centroids)
A, BC (with A & C centroids)
ABC (with B centroid)
Limited resolution

OTU groupings ignore the fine structure present in phylogeny
Same species, different genomes

Perna et al 2001 *Nature*, Welch et al 2002 *PNAS*

Three genomes, same species only 40% genes in common
Phylogeny: an alternative path

Many ecological analyses can be based on phylogeny:
- Alpha diversity (e.g. species diversity)
- Beta diversity (e.g. comparison of species across samples)
- Community assembly

So... what is a phylogeny, anyway?
Imagine you are dating a paleontologist...

VS.

T._rex  Stego  Veloci  Fluffy

T._rex  Stego  Veloci  Fluffy
Now imagine you've got dino DNA...

Let's try to reject the reviewer's phylogeny using DNA evidence!

Multiple alignment (MUSCLE, FSA, etc)

T_rex
ACC
>Stego
TCC
>Veloci
ACG
>Fluffy
ATCG
How does DNA evolve?

- Simplest model: all nucleotides are equally common, all changes from one to another equally likely (Jukes and Cantor, 1969)

Rate of substitution is $u/3$ per unit time

Expected number of changes on branch of length $t$ is $(4/3)ut$

Prob. of no change: $e^{-(4/3)ut}$

Prob. of at least one change: $1 - e^{-(4/3)ut}$

Prob. of e.g. A to C is $\text{Prob}(C|A,u,t) = (1/4)(1 - e^{-(4/3)ut})$
Calculating the likelihood of data given a tree

**Steps:**

1) Branch lengths
2) Finite-time transition probabilities
3) Leaf node partial probabilities

$$P(X|Y,u,t) = \frac{1}{4}(1 - e^{-(4/3)ut})$$

- $P(X|Y,0.1,1.0) = 0.0312$
- $P(X|X,0.1,1.0) = 0.9064$
- $P(X|Y,0.1,2.0) = 0.0585$
- $P(X|X,0.1,2.0) = 0.8244$

**q r s**

Finite time transition matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.91</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>C</td>
<td>0.03</td>
<td>0.91</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>G</td>
<td>0.03</td>
<td>0.03</td>
<td>0.91</td>
<td>0.03</td>
</tr>
<tr>
<td>T</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Calculating the likelihood of data given a tree

Sites evolve independently. Calculate site likelihoods one-at-a-time

\[
\begin{align*}
\text{T}_\text{rex} & \quad \text{A-CC} \\
\text{Stego} & \quad \text{-TCC}
\end{align*}
\]

Matrix multiply site T_rex 1:

\[
\begin{bmatrix}
\text{A} & \text{C} & \text{G} & \text{T}
\end{bmatrix}
\begin{bmatrix}
\text{A} & \text{C} & \text{G} & \text{T}
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.91 & 0.03 & 0.03 & 0.03 \\
0.03 & 0.91 & 0.03 & 0.03 \\
0.03 & 0.03 & 0.91 & 0.03 \\
0.03 & 0.03 & 0.03 & 0.91 \\
\end{bmatrix}
\begin{bmatrix}
1 & 0 & 0 & 0
\end{bmatrix}
\]

\[
= \begin{bmatrix}
0.91 \\
0.03 \\
0.03 \\
0.03 \\
\end{bmatrix}
\]

Matrix multiply site Stego 1:

\[
\begin{bmatrix}
\text{A} & \text{C} & \text{G} & \text{T}
\end{bmatrix}
\begin{bmatrix}
\text{A} & \text{C} & \text{G} & \text{T}
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.91 & 0.03 & 0.03 & 0.03 \\
0.03 & 0.91 & 0.03 & 0.03 \\
0.03 & 0.03 & 0.91 & 0.03 \\
0.03 & 0.03 & 0.03 & 0.91 \\
\end{bmatrix}
\begin{bmatrix}
1 & 1 & 1 & 1
\end{bmatrix}
\]

\[
= \begin{bmatrix}
0.91 \\
0.03 \\
0.03 \\
0.03 \\
\end{bmatrix}
\]

Joint prob T_rex & Stego:

\[
\begin{align*}
\text{T}_\text{rex} & \quad \text{A} .91 \\
\text{Stego} & \quad \text{A} .91
\end{align*}
\]
Calculating the likelihood of data given a tree

Tree likelihood is product of sites:
\[ L = 0.0007216 \]
\[ \log(L) = -9.536 \]
Hypothesis testing with tree likelihoods

The likelihood ratio test

\[ L = 0.00007216 \]

\[ L = 0.000010348 \]

Take the ratio of likelihoods:
\[ \frac{0.00007216}{0.000010348} = 6.973328 \]

Reviewer's tree \(~7\) times less likely
What if you don't know the tree?
Many methods for tree inference

- Parsimony, Distance, **Maximum Likelihood, Bayesian**
- Maximum Likelihood
  - FastTree, RAxML, GARLI, PHYML, etc.
- Bayesian
  - MrBayes, BEAST, PhyloBayes
  - All based on Markov chain Monte Carlo (MCMC) algorithms

Number of unrooted tree topologies with \( n \) tips:

\[
(2n - 3)!! = \frac{(2n - 3)!}{2^{n-2}(n-2)!}
\]

<table>
<thead>
<tr>
<th>Trees with:</th>
<th>4 tips</th>
<th>6 tips</th>
<th>8 tips</th>
<th>10 tips</th>
<th>50 tips</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>105</td>
<td>20,395</td>
<td>2,027,025</td>
<td>2.84 \times 10^{74}</td>
</tr>
</tbody>
</table>

Bottom line: tree inference is *hard*  
Estimated number of atoms in observable universe: \( \sim 10^{80} \)
Using phylogenies for microbial ecology

- Building phylogeny from >1M sequences: **impossible**
- Alternative: place new sequences on reference tree
  - RAxML-EPA: Berger *et al* 2011 *Systematic biology*
  - pplacer: Matsen *et al* 2010 *BMC Bioinformatics*

A “fat” tree: showing number of placements on each branch

Reference sequences

New seqs placed on tree
Handling uncertainty

- Bayesian placement (pplacer)
  - Calculate probability of new sequence on each branch
  - pplacer can do this quickly, analytically (no MCMC)

A single sequence with uncertain placement

Placement distribution viewed as a fat tree

Placement is starting to look better than OTUs
Uncertainty in many sequences

- Combine placement distributions from all seqs in sample
Using a placement distrib.: alpha diversity

• Phylogenetic diversity is sum length of branches covered

\[
\text{Sample PD is } 0.01 + 0.01 + 0.01 + 0.01 + 0.01 = 0.05
\]

• BWPD: Balance-weighted phylogenetic diversity (Barker 2002)
  – Intuition: weight the contribution each lineage makes to PD by its relative abundance
  – Weights can reflect placement uncertainty
BWPD_θ: partial weighting for PD

- A 1-parameter function interpolates between PD and BWPD (Matsen & McCoy 2013, *PeerJ*).
- When θ = 0 it is simply PD. θ = 1 it is BWPD.
- Matsen & McCoy compare:
  - OTU-based diversity metrics
  - Phylogenetic diversity (Faith 1992)
  - Phylogenetic quadratic entropy (Allen, Kon & Bar-Yam 2009)
  - qD(T) (Chao, Chiu, Jost 2010)
  - BWPD (Barker 2002)
  - BWPD_θ

on 3 different microbial communities, measuring correlation of diversity & phenotype
  - Vaginal, oral, & skin microbiomes

- θ=0.25 & θ=0.5 have highest correlation with microbial community phenotypes
- OTU based diversity metrics have least correlation with phenotype
Beta diversity: Edge Principal Component Analysis

• Edge PCA for exploratory data analysis (Matsen and Evans 2013)

• Given $E$ edges and $S$ samples:
  - For each edge, calculate difference in placement mass on either side of edge
  - Results in $E \times S$ matrix
  - Calculate $E \times E$ covariance matrix
  - Calculate eigenvectors, eigenvalues of covariance matrix

• Eigenvector: each value indicates how “important” an edge is in explaining differences among the $S$ samples

Example calculating a matrix entry for an edge:
This edge gets 5-2=3
Branches are thickened & colored according to the amount they shift the sample along an axis

Matsen & Evans 2012 *PLoS ONE*
Edge PCA and the vagina

- Samples colored according to Nugent score of bacterial vaginosis: blue → healthy, red → sick  
  
  (Matsen & Evans 2012)
How to do it?

1. Find reference sequences
2. Align reference sequences
3. Infer reference phylogeny
4. For each sample:
   4.1. Add sequences to alignment
   4.2. Place sequences on tree
5. Alpha & Beta diversity analysis

Each step is a unix command
PhyloSift: genome and metagenome phylogeny

Illumina reads placed onto reference gene family trees

- 40 “elite” families: universal among ~4000 Bact, Arch, Euk genomes (Lang et al 2013, Wu et al 2013)
- 350,000 “extended” families: SFAMs (Sharpton et al 2012)
- Amino-acid and nucleotide alignments+phylogenies

Darling et al 2014 PeerJ.
Using phylosift

Download phylosift: phylosift.wordpress.org

bin/phylosift all --output=hmp tutorial_data/HMP_1.fastq.gz

open hmp/HMP_1.fastq.gz.html

Shows taxonomic plot (Mac)

bin/guppy fpd --theta 0.25,0.5 hmp/*.gz.jplace

Alpha diversity

bin/guppy e pca --prefix pca hmp/*.gz.jplace

Beta diversity (min 3 samples)

More examples at: phylosift.wordpress.org

Raw illumina data
QIIME vs. PhyloSift

Data from Yatsunenko et al 2012. 16S amplicon & metagenomes from same samples.

Phylosift on proteins & 16S produces similar results to QIIME on amplicon data.
Phylogenetic alpha diversity

Data from Yatsunenko et al 2012

- Growth in PD over life
- BWPD is biphasic
PhyloSift compute requirements

- You don't need a huge computer to run PhyloSift
phylosift and major life events
On December 3$^{rd}$ 2010, Kai and his microbiome were born
Lots of nappies, lots of sampling

Kai Darling born 3rd Dec. 2010 in California, flew to Sydney 3.5 weeks later

March 1st 2011: a lot of poop in tubes and no idea how to get it through USA quarantine

Tiffanie Nelson at UNSW:
Extracted DNA with PowerSoil kits, mailed to USA
Metagenomics on a shoestring budget*

“Homebrew” Illumina Nextera library prep protocol:

**Goal:** metagenomics as easy as 16S amplicon studies

**Strategy:** Transposon-catalyzed library prep. 
Express & purify Tn5 from pWH1891. Custom adapters. 2.5ng input
Pool samples as early as possible.

**Results:** Sequenced 45 time points in HiSeq 2000 lane
~ $1 / library reagent costs, 100s of libraries in a day, NO ROBOTS

*Infant microbiome sequencing sponsored by private funds
PhyloSift view of fecal microbiome at three weeks age

- Tree-browsing of read placement mass (via archaeopteryx)
- Taxonomic summary plots in Krona (Ondov et al 2011)
Alpha diversity of gut communities vs. time

- Standard & balance-weighted PD (McCoy & Matsen, 2013)
- Phylogenetic diversity (PD) decreases?!

**Pearson's cor:** -0.44, \( p = 0.005 \)  
(p < 10\(^{-6}\) without formula samples)

**Pearson's cor:** 0.21, \( p = 0.18 \)  
(\( p = 0.0071 \) w/o formula)

*B. thetaiotaomicron* becomes dominant
Phylogenetic “Edge PCA” on infant fecal microbiome

Edge PCA: explain variation in community structure among many samples

Matsen & Evans 2013 PLoS ONE

Infant gut timeseries

Up: Bifidobacterium longum

3rd PC (1%):
Staphylococcus Veillonella

Up: Bacteroides, Down: Bifidobacterium
Formula-fed samples within one day

Up: Bifidobacterium longum

A day on formula

One week on formula,
Six poops in one day.

Up: Bacteroides, Down: Bifidobacterium
Thanks!