Phylogenetic Geometry

Ruth Davidson

University of Illinois Urbana-Champaign
Department of Mathematics

Mathematics and Statistics Seminar
Washington State University-Vancouver
September 26, 2016
Goal of phylogenetics: mine **genetic** and **genomic** data...

- nucleotide or protein sequences
- morphology
- behavioral traits

for phylogenetic signal to turn into phylogenies!
Example Pipeline: Sample Data to Phylogeny

1. Sample from extant taxa:

2. Align samples:

   A G G G C A T
   T A G C C C A
   T A G A C T T
   A G C A C A A
   A G C G C T T

3. Compute a measure of dissimilarity, such as:

   Hamming(AGCACAA, AGCGCTT) = 3

4. Dissimilarity measure ⇒ your favorite method ⇒ phylogeny!
Reality: genomes are much longer than 7 nucleotides, want trees on more than 5 taxa

Challenge: need statistical models and methods to find species phylogenies on “big data”: samples across many regions of many genomes
Example: the Jukes-Cantor (JC69) Model of Sequence Evolution

- JC69 is a time-reversible, continuous time, stationary Markov process.
- JC69 is defined by a rate matrix $Q$, a probability matrix (for each tree branch) $P(t) = \exp(Qt)$, e.g. $P'(t) = Q \cdot P(t)$

**Base Frequencies:**
\[ \pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4} \]

**Rate Matrix $Q$:**

<table>
<thead>
<tr>
<th></th>
<th>$A$</th>
<th>$G$</th>
<th>$C$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$-\frac{3}{4}\mu$</td>
<td>$\frac{\mu}{4}$</td>
<td>$\frac{\mu}{4}$</td>
<td>$\frac{\mu}{4}$</td>
</tr>
<tr>
<td>$G$</td>
<td>$\frac{\mu}{4}$</td>
<td>$-\frac{3\mu}{4}$</td>
<td>$\frac{\mu}{4}$</td>
<td>$\frac{\mu}{4}$</td>
</tr>
<tr>
<td>$C$</td>
<td>$\frac{\mu}{4}$</td>
<td>$\frac{\mu}{4}$</td>
<td>$-\frac{3\mu}{4}$</td>
<td>$\frac{\mu}{4}$</td>
</tr>
<tr>
<td>$T$</td>
<td>$\frac{\mu}{4}$</td>
<td>$\frac{\mu}{4}$</td>
<td>$\frac{\mu}{4}$</td>
<td>$-\frac{3\mu}{4}$</td>
</tr>
</tbody>
</table>

**Example.** $\Pr(\text{observing state } A \text{ for taxon 1 } | \text{ state } G \text{ at } u) =$

\[ \pi_G \cdot P_{A,G}(t) = \left( \frac{1}{4} \right) \left( \frac{1}{4} \left( 1 - e^{-t\mu} \right) \right) \]

**Example.** JC69 evolutionary distance in expected number of changes:

\[ \text{JC69}(\text{AGCACA}A, \text{AGCGC}T)T = -\frac{3}{4} \ln \left( 1 - \frac{4}{3} \left( \frac{3}{7} \right) \right) \approx 0.635 \]
Examples of Methods Phylogeneticists Use

For estimating gene trees

- Distance-based methods (UPGMA, NJ, BME, LSP, FastME, BIONJ)
- Maximum-likelihood methods (RAxML, FastTree, PhyML)
- Parsimony methods (MRP in PAUP*)

Methods for estimating genome (species) trees:

- Bayesian methods (MrBayes, BUCKy, *BEAST)
- Gene tree methods applied after concatenation of gene alignments
- Summary methods (NJst, ASTRAL-II, wQMC)
- Quartet-based methods (SVDquartets, QMC, Quartet-puzzling)

Methods for co-estimating alignments and genome phylogenies:

- Divide-and-conquer recursive methods (SATé-II, PASTA)
- Bayesian co-estimation methods (Bali-PHY)
Oh No! How Do I Choose or Design a Good Method!!!???

Evaluations of Method Performance:

- Algorithmic running time ($O(\log n)$, $O(n^4)$, ...)
- Empirically derived confidence
- Statistical consistency (no positively misleading cases OR more good data $\Rightarrow$ more accurate results...)
- Accuracy (under a measure of tree dissimilarity) on simulated and "curated" data...

**Phylogenetic Geometry**

- Novel methods for detection of accuracy and bias
- Novel scoring mechanisms for existing methods
Methods Subject to Geometric Perspectives

For estimating gene trees:

- Distance-based methods (UPGMA, NJ, BME, LSP, FastME, BIONJ)
- Maximum-likelihood methods (RAxML, FastTree, PhyML)
- Parsimony methods (MRP in PAUP*)

Methods for estimating genome (species) trees:

- Bayesian methods (MrBayes, BUCKy, *BEAST)
- Gene tree methods applied after concatenation of gene alignments
- Summary methods (NJst, ASTRAL-II, wQMC)
- Quartet-based methods (SVDquartets, QMC, Quartet-puzzling)

Methods for co-estimating alignments and genome phylogenies:

- Divide-and-conquer recursive methods (SATé-II, PASTA)
- Bayesian co-estimation methods (Bali-PHY)
This talk:

For estimating gene trees:

- **Distance-based methods** (UPGMA, NJ, BME, LSP, FastME, BIONJ)
- Maximum-likelihood methods (RAxML, FastTree, PhyML)
- Parsimony methods (MRP in PAUP*)

Methods for estimating genome (species) trees:

- Bayesian methods (MrBayes, BUCKy, *BEAST)
- Gene tree methods applied after concatenation of gene alignments
- **Summary methods** (NJst, ASTRAL-II, wQMC)
- **Quartet-based methods** (SVDquartets, QMC, Quartet-puzzling)

Methods for co-estimating alignments and genome phylogenies:

- Divide-and-conquer recursive methods (SATé-II, PASTA)
- Bayesian co-estimation methods (Bali-PHY)
Distance-based methods: inputs $D$ vs. outputs $d$

- Input: a *dissimilarity map* $D$ on taxa $[n] = \{1, 2, \ldots, n\}$ satisfying $D(x, y) = D(y, x)$ for all $\{x, y\} \subset [n]$, and $D(x, x) = 0$.

- Output: a *tree metric* $d$ on taxa $[n]$ comprised of pairwise path distances in a weighted tree.

\[
D = \begin{bmatrix}
1 & 2 & 3 \\
1 & 0 & 1 \\
2 & 1 & 0 \\
3 & 3 & 5 \\
\end{bmatrix}
\]

UPGMA $\rightarrow$

\[
D(1) = \begin{bmatrix}
1, 2 & 3 \\
1, 2 & 0 \\
1, 2 & 4 \\
3 & 4 \\
\end{bmatrix}
\]

\[
D = \begin{bmatrix}
1, 2 \\
1, 3 \\
2, 3 \\
(1, 3, 5) \\
\end{bmatrix}
\]

UPGMA $\rightarrow$

\[
d = \begin{bmatrix}
1, 2 & 1, 3 & 2, 3 \\
1, 2 & 3 & 4 \\
(1, 4, 5) & (1, 4, 5) & (1, 4, 5) \\
\end{bmatrix}
\]
UPGMA is a Heuristic for Least-Squares Phylogeny (LSP)

**Theorem (Day 1987).** LSP is NP-Hard.

\[
D = \begin{pmatrix}
1, 2 & 1, 3 & 2, 3 \\
1 & 3 & 5
\end{pmatrix}
\xrightarrow{UPGMA} d = \begin{pmatrix}
1, 2 & 1, 3 & 2, 3 \\
1 & 4 & 4
\end{pmatrix}
\]

Grey lines are the boundaries of the **UPGMA cones**
Analysis of UPGMA

- Algorithmic running time? $O(n^2)$
- Empirically derived confidence? No: overwhelming evidence of bias towards balanced trees (Aldous 2001)
- Statistical consistency? No:
- Accuracy as heuristic for LSP? Only in special cases (Molecular clock.....)

**Phylogenetic Geometry:**

*Theorem (D-Sullivant 2013).* The combinatorial structure of the UPGMA cones indicates that unbalanced trees have smaller cones.

*Theorem (D-Sullivant 2014).* The combinatorial structure of the UPGMA cones leads to overwhelming bias against unbalanced trees near certain points in $\mathbb{R}_{\geq 0}^{\binom{n}{2}}$. 
The Neighbor-Joining (NJ) Algorithm (Saitou and Nei 1987)

**NJ Algorithm:**

- **Input:** dissimilarity map $D \in \mathbb{R}^{(n\choose 2)}_{\geq 0}$.

- **Step 1:**
  - Linearly transform the input $D$ using the map
    $$q_{x,y} = (n - 2)D_{x,y} - \sum_{k=1}^{n} D_{x,k} - \sum_{k=1}^{n} D_{y,k}$$
  - Identify $(x, y)$ as a single node
  - Update distances to node $(x, y)$:
    $$D(1)_{(x,y),z} = (1/2)(D_{x,z} + D_{y,z} - D_{x,y})$$

- Recurse Step 1 on nodes (clusters) until
  $$D(r) \in \mathbb{R}^{(2\choose 2)}.$$  

- **Output:** arbitrary tree metric $d$ on $n$ leaves.

**Observation:** NJ divides $\mathbb{R}^{(n\choose 2)}_{\geq 0}$ into a family of polyhedral cones indexed by (1) tree shape and (2) the order of agglomeration of clusters of taxa.
The Balanced Minimum Evolution (BME) (Desper-Gascuel 2002)

**Definition (BME Criterion).**
Let $T$ be a tree. Let $\lambda_{x,y} = \prod_{v \in p^T_{x,y}} \left( \deg(v) - 1 \right)^{-1}$ if $x \neq y$, and $\lambda^T_{x,x} = 0$, $(p^T_{x,y} = \{\text{internal } v \in V(T) \text{ on path between leaves } x \text{ and } y\})$. The **BME Criterion** for a tree metric $d$ is

$$\sum_{(x,y):x \text{ and } y \text{ are leaves of } T} \lambda_{x,y} d_{x,y}.$$ 

**Theorem (Gascuel-Steel 2006).** NJ is a greedy heuristic for minimizing the BME criterion.

**Theorem (Pauplin 2000).** The BME criterion divides $\mathbb{R}^{\binom{n}{2}}$ into a family of cones indexed by the optimal solutions to BME.

**Definition.** The **BME polytope** is the convex hull of the vectors

$$\{\lambda_{(1,2)^T}, \lambda_{(1,2)^T}, \ldots, \lambda_{(n-1,n)^T} : T \text{ is a tree with } n \text{ leaves}\}.$$ 

**Theorem (Pachter-Sturmfels 2005).** Finding the BME tree is equivalent to optimizing a linear functional over the BME polytope.
Analysis of NJ

- Algorithmic running time? \( O(n^3) \)

- Statistical consistency (re: LSP)? Yes:

\[
\begin{align*}
1 & \quad d \\
& \quad w \\
& \quad 3 \quad 4 \\
\end{align*}
\]

**Theorem (Atteson 1999).** If \( \|D - d\|_\infty < w/2 \), NJ \( (D) = \hat{d} \), and \( \hat{d} \sim d \).

- Accuracy as heuristic for LSP? Great! (see above) except near certain points in \( \mathbb{R}^{2 \choose 2}_{\geq 0} \) (D-Sullivant 2014).

- Phylogenetic Geometry:

**Theorem (Eickmeyer-Huggins-Pachter-Yoshida 2008).** Numerical study (for taxa \( \leq 8 \)) of NJ cones and BME indicates accuracy as a heuristic for BME.

**Theorem (Haws-Hodge-Yoshida 2011).** The combinatorial structures of the families of (1) NJ cones and (2) BME cones guarantees accuracy in key cases as a heuristic for BME.
For estimating gene trees:

- Distance-based methods (UPGMA, NJ, BME, LSP, FastME, BIONJ)
- Maximum-likelihood methods (RAxML, FastTree, PhyML)
- Parsimony methods (MRP in PAUP*)

Methods for estimating genome (species) trees:

- Bayesian methods (MrBayes, BUCKy, *BEAST)
- Gene tree methods applied after concatenation of gene alignments
- **Summary methods** (NJst, ASTRAL-II, wQMC)
- **Quartet-based methods** (SVDquartets, QMC, Quartet-puzzling)

Methods for co-estimating alignments and genome phylogenies:

- Divide-and-conquer recursive methods (SATé-II, PASTA)
- Bayesian co-estimation methods (Bali-PHY)
Multispecies Coalescent & Gene Tree Discordance

Gene tree = species tree

Incomplete Lineage Sorting (ILS)

Horizontal Gene Transfer (HGT)

Gene Tree Invariants for Quartets

\[ T = \text{model tree with unrooted shape } \left( (a, b), c, d \right), \ p_i = \Pr\{q_i|T\} \]

**Theorem (Allman-Degnan-Rhodes 2011).** Under the MSC, \( p_1 > p_2 = p_3 \).

- The polynomial \( f(x) = x_2 - x_3 \) is a **phylogenetic invariant**: \( f(p_2, p_3) = 0 \).
- Methods using invariants of site pattern probabilities: (Examples of quartet-based! ) [Casanellas & Fernandez-Sanchez 2006, Rusinko-Hipp 2012, Chifman-Kubatko 2014 ... ]
- Sadly, though many improvements made, no invariant-based methods have decisively outperformed all other methods!
Hope for Gene Tree Invariants: Species Trees as Optimization Problems

**Theorem (Allman-Degnan-Rhodes 2011).** $p_1 > p_2 = p_3$ under the MSC.

**Theorem (Roch, Snir 2013).** $p_1 > p_2$ and $p_1 > p_3$ for bounded levels of HGT (under two models of HGT).

**Problem (Maximum Quartet Support Species Tree Problem).**
- Input: set $\mathcal{T}$ of unrooted gene trees on species set $S$.
- Output: species tree $T$ maximizing $\sum_{q \in Q(T)} w(q, T)$:
  
  $Q(T) =$ set of quartet trees induced by $T$,
  
  $w(q, T) = |\{t \in \mathcal{T} : q \in Q(t)\}|$

If we include inequalities in our definition of invariants, MQSST is an invariant-based inference problem!

**Theorem (D et. al. 2015).** The polynomial-time summary method ASTRAL-II, which solves MQSST, is statistically consistent under the MSC model and the bounded HGT of Roch-Snir 2013.

**Conjecture (Comparative study in D et. al. 2015).** Inequality-invariant based summary methods will perform competitively in the face of multiple sources of gene tree discordance.
The Role of Quartet Agglomeration Methods

For estimating gene trees:
- Distance-based methods (UPGMA, NJ, BME, LSP, FastME, BIONJ)
- Maximum-likelihood methods (RAxML, FastTree, PhyML)
- Parsimony methods (MRP in PAUP*)

Methods for estimating genome (species) trees:
- Bayesian methods (MrBayes, BUCKy, *BEAST)
- Gene tree methods applied after concatenation of gene alignments
- Summary methods (NJst, ASTRAL-II, wQMC)
- **Quartet-based methods** (SVDquartets, QMC, Quartet-puzzling)

Methods for co-estimating alignments and genome phylogenies:
- Divide-and-conquer recursive methods (SATé-II, PASTA)
- Bayesian co-estimation methods (Bali-PHY)
Theorem (Roch-Warnow 2015). Gene tree estimation error cannot be bounded in general.

Problem. Gene tree estimation is a necessary component of summary method pipelines

- SVDquartets (Chifman-Kubatko 2014) scores quartets by measuring the Frobenius norm distance to algebraic varieties (Chifman-Kubatko 2015) encoded by matrices of site pattern probabilities for sequences of length 1 on 4 leaves:

\[
\begin{pmatrix}
    p_{AAAA} & p_{AACA} & \cdots & p_{ATTA} \\
    p_{AAAC} & p_{AACC} & \cdots & p_{ATTC} \\
    \vdots & \vdots & \ddots & \vdots \\
    p_{TAAT} & p_{TACT} & \cdots & p_{TTTT}
\end{pmatrix}
\]

- Then, the best-scoring species-level quartets are fed to a quartet-agglomeration method such as QMC (Snir-Rao 2012) or QFM (Reaz et al. 2014), also PAUP*-beta (Swofford) to find the species tree
SVDquartets Pipeline and Performance Issues

- Running time: slow!
- Accuracy: Excellent on JC69-simulated sequences
- Empirically Derived confidence: Mixed to Promising (Chou et. al 2015, Leavitt et al. 2016)
- Statistical Consistency: impossible to separate from quartet-agglomeration methods QFM and QMC, for which this cannot be established

**Problem.** Summarizing quartets from 50-taxon datasets is still computationally prohibitive.

**Solution:** Reduce the number of quartets needed for (QMC, QFM, Quartet-puzzling...) to return a species phylogeny with small loss in accuracy.

**Theorem.** (Davidson-Lawhorn-Rusinko-Weber 2016+) There exists an "efficient quartet system" (EQS) that definitively represents a tree structure.

See [www.github.com/redavids/efficientquartets](http://www.github.com/redavids/efficientquartets)
Thanks to WSU Vancouver for Your Hospitality


- To you for listening!
- To my many collaborators!
- To my recent funding: NSF DMS-1401591, NSF DMS-0954865,

www.math.uiuc.edu/~redavid2/