A Signal-to-Noise Analysis of Phylogeny Estimation by Neighbor-Joining

Insufficiency of Polynomial Length Sequences

Michelle R. Lacey[1] Joseph. T. Chang[2]

[1]Department of Mathematics
Tulane University

[2]Department of Statistics
Yale University

Future Directions in Phylogenetic Methods and Models, 2007

Why study Neighbor-Joining?

NJ has nice properties:

- Consistency
- Computational efficiency
- Strong performance in numerous simulation studies

". . . NJ should be regarded as a universal lowest common denominator in testing reconstruction algorithms. Its speed makes it easy to use under all circumstances; its topological accuracy makes it an acceptable starting point for tree reconstruction in biological practice. Thus any new proposed method should first be compared to NJ and abandoned if it does not offer an advantage over NJ for at least substantial subsets of the parameter space." – K. St. John, T. Warnow, B.M.E. Moret, L. Vawter, *Journal of Algorithms* (2003)

The Neighbor-Joining Algorithm

Let $\{d_{ij}\}$ be the set of pairwise distances for a set of n sequences.

1. From the set of n leaf nodes, select the pair i, j that minimizes the criterion $D_{ij} = d_{ij} - \frac{1}{n-2}\left(\sum_{k=1}^{n} d_{ik} + \sum_{k=1}^{n} d_{jk}\right)$
2. Define a new node k and set $d_{km} = \frac{1}{2}\left(d_{im} + d_{jm} - d_{ij}\right)$ for all leaves m.
3. Set $d_{ik} = \frac{1}{2}\left(d_{ij} + \frac{1}{n-2}\left(\sum_{m=1}^{n} d_{im} - \sum_{m=1}^{n} d_{jm}\right)\right)$ and $d_{jk} = d_{ij} - d_{ik}$.
4. Remove i and j from the set of active nodes and add k.
5. Repeat steps 1-4 until only two nodes i^* and j^* remain, and connect these with an edge of length $d_{i^*j^*}$.

Convergence results for N-J

Kevin Atteson, *Algorithmica* (1999)

Under the Cavender-Farris model, let E be the set of edges in a tree T with n leaves, and let d_e be the length of a given edge $e \in E$. Let $\epsilon \leq \min(d_e)/2$. Then the N-J algorithm will recover the true tree topology with probability at least $1 - \delta$ if

$$k \geq \frac{8 \ln\left(n^2/\delta\right)}{\left(1 - \exp(-\epsilon)\right)^2} \exp\left(\max_{i,j} 4 d_{ij}\right)$$

Interpretation: N-J will, with high probability, accurately reconstruct any tree with n leaves from sequences that are exponential in the length of the longest pairwise distance d_{ij}.

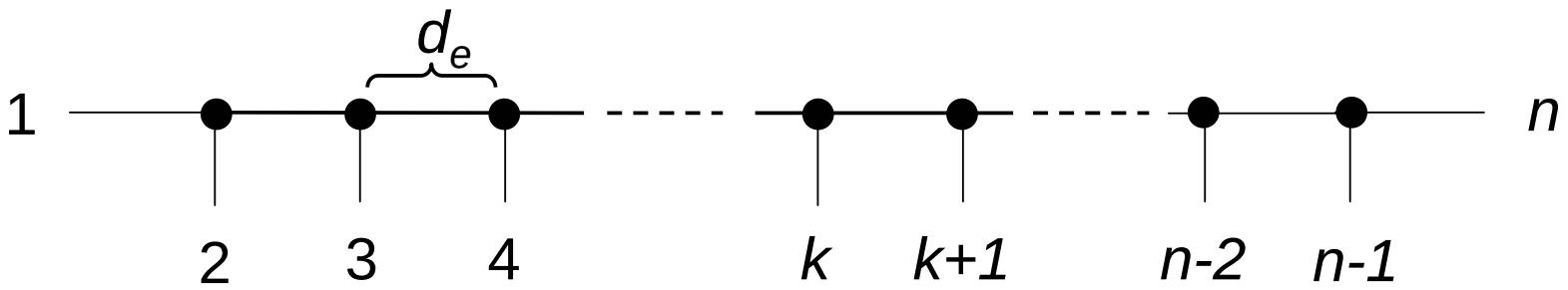
Exponential convergence is good, but. . .

Fast-convergence: A phylogeny reconstruction method is defined to be *fast-converging* under a model of evolution if the method can, with high probability, accurately recover the topology of any model tree from sequences that grow only polynomially in the number of leaves. (Huson, Nettles, and Warnow, *Journal of Computational Biology* (1999))

Central question: Is NJ a fast-converging method?

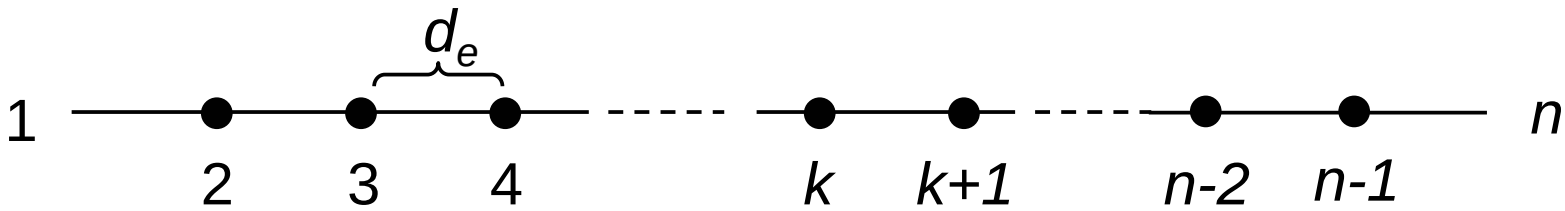
Case study: The Caterpillar Tree

The term “caterpillar tree” is used to describe a phylogeny in which *n* taxa are connected to a single spine.



Case study: The Caterpillar Tree

If we consider a simplified "legless" caterpillar in which the n taxa are connected in sequence by edges of equal length d_e, then the distance between a pair of taxa i and j on a caterpillar tree is simply given by $|j - i|d_e$, the number of edges separating the pair.

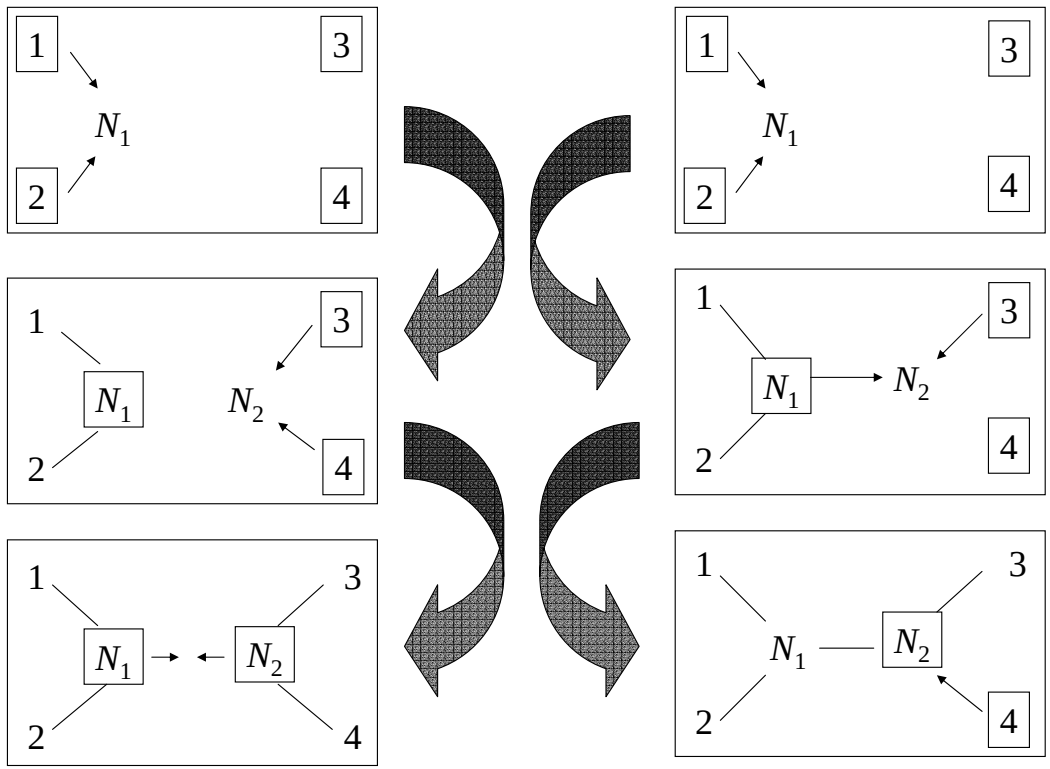


Why study this tree? Since the longest pairwise distance $d_{1,n} = (n - 1)d_e$, Atteson's bound would require sequence lengths to be exponential in n to guarantee asympototic convergence.

Reconstructing a caterpillar

A caterpillar tree can only be correctly reconstructed by working from the outside in, joining either leaves 1 and 2 or leaves $n-1$ and n on the first step.

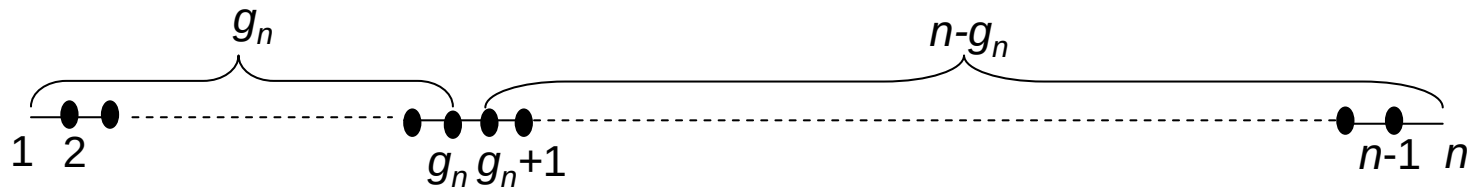
Two of the four correct ways to reconstruct a 4-leaf caterpillar tree:



Analytical objective

We've just seen that the NJ criterion must be minimized by either $\hat{D}_{12}$ or $\hat{D}_{n-1,n}$ for the algorithm to correctly join two neighboring leaves on the first step.

Suppose we consider another pair of sequences, S_{g_n} and S_{g_n+1}, with $g_n > 2$, where $g_n = n^\gamma$ for any $\gamma \in \left(0, \frac{1}{2}\right)$



Define the variable $D_n = \left(\hat{D}_{g_n,g_n+1} - \hat{D}_{12}\right)$. A necessary (but clearly insufficient) condition for fast-convergence is that $P(D_n > 0) \approx 1$ for sequences of polynomial length in n. Thus, our mission is to understand the asymptotic behavior of D_n.

But there's a slight complication...

- Under the Cavender-Farris model, $\hat{d}_{ij} = -\frac{1}{2} \ln \left(1 - 2\hat{p}_{ij}\right)$, which is undefined whenever $\hat{p}_{ij} \geq \frac{1}{2}$.
- For edges of length p_e, $p_{ij} = \frac{1}{2}\left(1 - (1 - 2p_e)^{j-i}\right)$
- For sequences of polynomial length $L_n = n^s$, $P(\hat{p}_{ij} > \frac{1}{2}) \to \frac{1}{2}$ for distant pairs i and j.

What to do? Well, we could stop right here, but instead we allow for the "correction" of undefined distances, setting $\hat{d}_{ij} = d^\star = \frac{1}{2} \ln \left(\frac{L_n}{2}\right)$, the maximum observable distance for sequences of length L_n.

Statistical approach

Instead of attempting to fully characterize the probability distribution of D_n, we derive two inequalities:

1. An asymptotic upper bound for the expectation $E(D_n)$
2. An asymptotic lower bound for the variance $Var(D_n)$

With these results, we analyze the convergence properties of the "signal-to-noise" ratio $\frac{E(D_n)}{\sqrt{Var(D_n)}}$.

Derivation of an upper bound for the expectation of D_n

Define $\hat{d}_{i.} = \sum_{k=1}^{n} \hat{d}_{i,k}$, and let $L_n = n^s$ for any $s > 1$.

$$
\begin{aligned}
E(D_n) &= E\left(\hat{d}_{g_n,g_n+1} - \frac{1}{n-2}\left(\hat{d}_{g_n.} + \hat{d}_{g_n+1.}\right) - \hat{d}_{12} + \frac{1}{n-2}\left(\hat{d}_{1.} + \hat{d}_{2.}\right)\right) \\
&= E\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{12}\right) + \frac{1}{n-2}\left(E\left(\hat{d}_{1.} - \hat{d}_{g_n.}\right) + E\left(\hat{d}_{2.} - \hat{d}_{g_n+1.}\right)\right) \\
&= \frac{1}{n-2}\left(E\left(\hat{d}_{1.} - \hat{d}_{g_n.}\right) + E\left(\hat{d}_{2.} - \hat{d}_{g_n+1.}\right)\right).
\end{aligned}
$$

Subdivide the summations into three regions:

- Region 1: $k \in [1, g_n + 1]$
- Region 2: $k \in [g_n + 2, g_n + b_n]$ for $b_n = n^{\beta}, \beta \in \left(0, \frac{1}{2}\right)$
- Region 3: $k > g_n + b_n$

Expectation results

- For Region 1, $\sum_{k=1}^{g_n+1} E\left(\left(\hat{d}_{1,k} - \hat{d}_{g_n k}\right) + \left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right)\right) = 0$
- In Region 2, distances $d_{1,k}$ are much larger than $d_{g_n,k}$, so we bound the expectation conservatively:

$$\sum_{k=g_n+2}^{g_n+b_n} E\left(\left(\hat{d}_{1,k} - \hat{d}_{g_n,k}\right) + \left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right)\right)$$

$$< \sum_{k=g_n+2}^{g_n+b_n} \left(\left(d^\star - E(\hat{d}_{g_n,k})\right) + \left(d^\star - E(\hat{d}_{g_n+1,k})\right)\right) < 2b_n d^\star < b_n s \ln(n)$$

- All distances are large in Region 3, and for sequences of length n^s we find that the expectation is negligible:

$$\sum_{k=g_n+b_n+1}^{n} E\left(\left(\hat{d}_{1,k} - \hat{d}_{g_n,k}\right) + \left(\hat{d}_{2,k} - \hat{d}_{g_n+1,k}\right)\right) < \frac{\ln(n)}{p_e} s n^s (1-2p_e)^{b_n}$$

Upper Bound for $E(D_n)$

Aggregating these results, an overall upper bound is given by

$$E(\hat{D}_{g_n,g_n+1} - \hat{D}_{12}) \leq \frac{\ln(n)}{n-2}\left(sb_n + \frac{sn^s(1-2p_e)^{b_n}}{p_e}\right)$$
$$= \frac{\ln(n)}{n-2}\left(sn^{\beta} + o(n^{-1})\right)$$

Derivation of a lower bound for the variance of D_n

$$\mathrm{Var}(D_n) = \mathrm{Var}\left[\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{1,2}\right) + \frac{1}{n-2}\left(\left(\hat{d}_{1.} - \hat{d}_{g_n.}\right) + \left(\hat{d}_{2.} - \hat{d}_{g_n+1.}\right)\right)\right]$$

Bounding this expression makes use of the following results:

- $d_{1,2} = d_e = d_{g_n,g_n+1}$ can be estimated with great precision by sequences of length n^s, so variance and covariance terms involving $\left(\hat{d}_{g_n,g_n+1} - \hat{d}_{1,2}\right)$ converge to 0.
- For $k > g_n$ and for n sequences of length $L = n^s$ for any fixed s, if $g_n = n^\gamma$ for any $\gamma \in \left(0, \frac{1}{2}\right)$, then $\mathrm{Var}(\hat{d}_{1,k} - \hat{d}_{g_n,k}) \geq \left(\frac{1}{4} - \frac{\delta_{g_n,k}}{2} - o(n^{-1})\right)\left(\frac{s}{4} - \frac{1}{2} - \frac{s}{n}\right)^2 (\ln(n))^2$ with $\delta_{g_n,k} = P\left(\hat{p}_{g_n,k} < \frac{1}{2}\right) - \frac{1}{2}$.
- For all sequences S_i, S_j, S_k, and S_l, $\mathrm{Cov}\left(\hat{d}_{i,j}, \hat{d}_{k,l}\right) \geq 0$

Lower Bound for $Var(D_n)$

After lots of calculations, we find

$$\mathrm{Var}(D_n) \geq \left(\frac{\ln(n)}{n-2}\right)^2 \left[2\left(n-\left(n^{\gamma}+n^{\beta}\right)\right)c_{\beta,s,n} - s^2 n^{2\gamma} - o\left(n^{-1}\right)\right]$$

where $c_{\beta,s,n} = \left(\frac{1}{4} - \frac{\delta_{g_n,k}}{2} - o(n^{-1})\right)\left(\frac{s}{4} - \frac{1}{2} - \frac{s}{n}\right)^2 \to \frac{1}{4}\left(\frac{s}{4} - \frac{1}{2}\right)^2$ as $n \to \infty$

Signal-to-Noise Ratio

For $\beta, \gamma \in \left(0, \frac{1}{2}\right)$ and $L_n = n^s$, we have

- $E(D_n) \leq \frac{\ln(n)}{n-2}\left(sn^{\beta} + o\left(n^{-1}\right)\right)$
- $\mathrm{Var}(D_n) \geq \left(\frac{\ln(n)}{n-2}\right)^2 \left[2\left(n - \left(n^{\gamma} + n^{\beta}\right)\right) c_{\beta,s,n} - s^2 n^{2\gamma} - o\left(n^{-1}\right)\right].$

Take the ratio (ignoring constants and those terms which converge to 0):

$$\begin{aligned} \frac{E(D_n)}{\mathrm{SD}(D_n)} &\leq \frac{n^{\beta}}{\sqrt{n - \left(n^{\gamma} + n^{\beta} + n^{2\gamma}\right)}} \\ &= \frac{1}{n^{\frac{1}{2}-\beta}\sqrt{1 - \left(n^{-(1-\gamma)} + n^{-(1-\beta)} + n^{-(1-2\gamma)}\right)}}. \end{aligned}$$

Implications

- For any $\beta, \gamma \in \left(0, \frac{1}{2}\right)$, $\lim_{n\to\infty} \frac{E(D_n)}{\mathrm{SD}(D_n)} = 0$, so the variability of D_n increases much more rapidly than its expected value.
- If we assume a few mild conditions on the distribution of D_n, then the result implies that

$$\lim_{n\to\infty} P(D_n < 0) = \lim_{n\to\infty} P\left(\frac{D_n - E(D_n)}{\sigma_{D_n}} < \frac{-E(D_n)}{\sigma_{D_n}}\right) = \frac{1}{2}.$$

Closing Remarks

- Exponential bound cannot be improved for general trees
- Result is not necessarily a criticism of NJ in practice, since the caterpillar tree is obviously an extreme case
- High variability caused by large number of poorly estimated distances
- NJ can be especially susceptible to such errors due to the nature of the algorithm

Thanks for listening! For more details, please see our paper: Lacey,MR and Chang, JT, *Mathematical Biosciences* 2006; 199(2): 188-215.