The Generalized Neighbor Joining method

Ruriko Yoshida
Dept. of Statistics University of Kentucky

Joint work with Dan Levy and Lior Pachter

`www.math.duke.edu/~ruriko`

Challenge

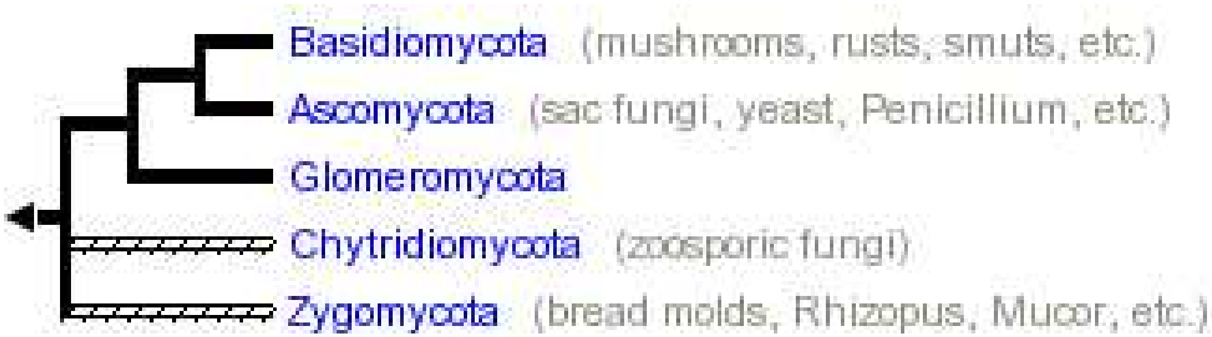
We would like to assemble the fungi tree of life.

Francois Lutzoni and Rytas Vilgalys Department of Biology, Duke University

$1500+$ fungal species

http://ocid.nacse.org/research/aftol/about.php

Many problems to be solved....

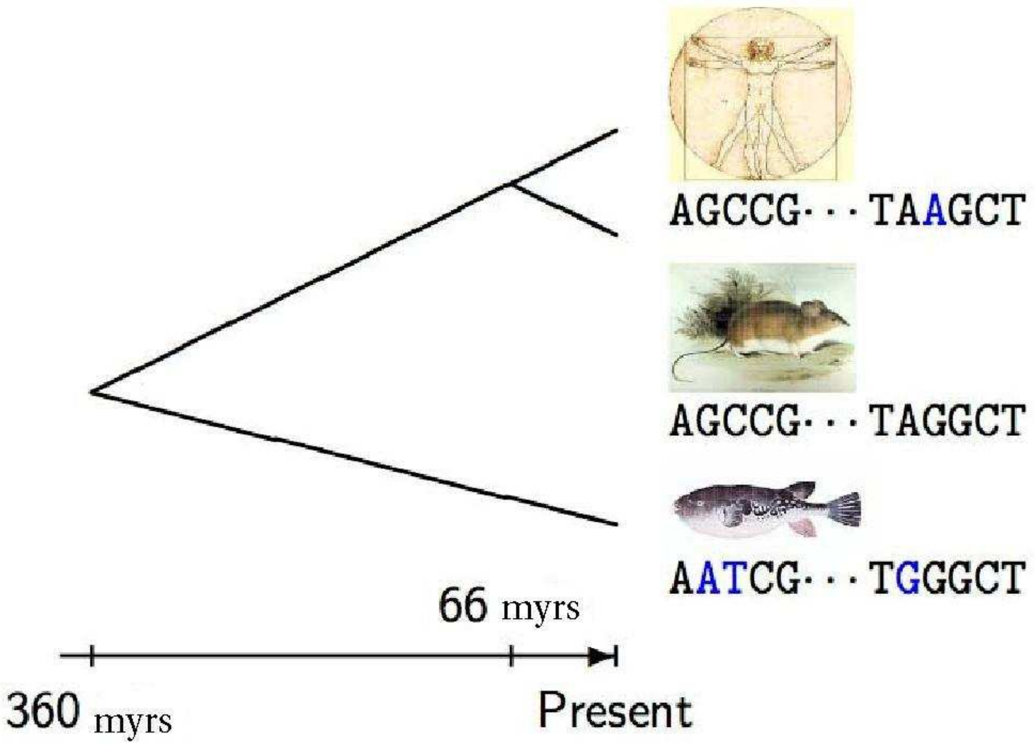

Basidiomycota (mushrooms, rusts, smuts, etc.)
Ascomycota (sac fungi, yeast, Penicillium, etc.)
Glomeromycota
Chytridiomycota (zoosporic fungi)
Zygomycota (bread molds, Rhizopus, Mucor, etc.)

`http://tolweb.org/tree?group=fungi`

Zygomycota is not monophyletic. The position of some lineages such as that of Glomales and of Engodonales-Mortierellales is unclear, but they may lie outside Zygomycota as independent lineages basal to the Ascomycota-Basidiomycota lineage (Bruns et al., 1993).

Phylogeny

Phylogenetic trees describe the evolutionary relations among groups of organisms.



Constructing trees from sequence data

“Ten years ago most biologists would have agreed that all organisms evolved from a single ancestral cell that lived 3.5 billion or more years ago. More recent results, however, indicate that this family tree of life is far more complicated than was believed and may not have had a single root at all.” (W. Ford Doolittle, (June 2000) *Scientific American*).

Since the proliferation of Darwinian evolutionary biology, many scientists have sought a coherent explanation from the evolution of life and have tried to reconstruct phylogenetic trees.

Methods to reconstruct a phylogenetic tree from DNA sequences include:

- **The maximum likelihood estimation (MLE) methods**: They describe evolution in terms of a discrete-state continuous-time Markov process. The substitution rate matrix can be estimated using the **expectation maximization (EM) algorithm**. (for eg. Dempster, Laird, and Rubin (1977), Felsenstein (1981)).

- **Distance based methods**: It computes pair-wise distances, which can be obtained easily, and combinatorially reconstructs a tree. The most popular method is the **neighbor-joining (NJ) method**. (for eg. Saito and Nei (1987), Studier and Keppler (1988)).

However

The MLE methods: An exhaustive search for the ML phylogenetic tree is computationally prohibitive for large data sets.

The NJ method: The NJ phylogenetic tree for large data sets loses so much sequence information.

Goal:

- Want an algorithm for phylogenetic tree reconstruction by combining the MLE method and the NJ method.
- Want to apply methods to very large datasets.

Note: An algebraic view of these discrete stat problems might help solve this problem.

The generalized neighbor-joining mathod

The GNJ method: in 2005, Levy, Y., and Pachter introduced the **generalized neighbor-joining (GNJ) method**, which reconstructs a phylogenetic tree based on comparisons of subtrees rather than pairwise distances

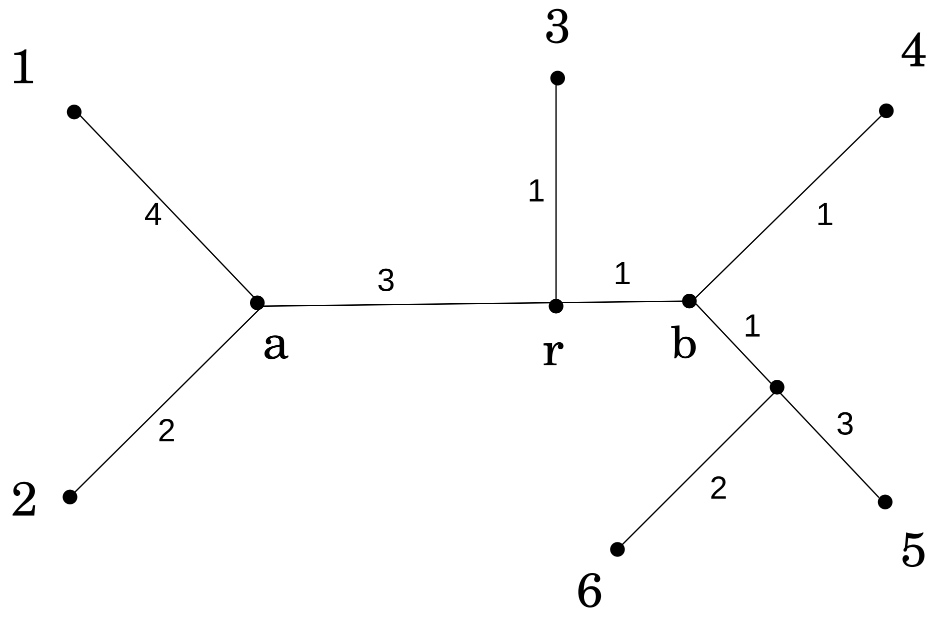
- The GNJ method is a method combined with the MLE method and the NJ method.
- The GNJ method uses more sequence information: the resulting tree should be more accurate than the NJ method.
- The computational time: polynomial in terms of the number of DNA sequences.

The GNJ method

MJOIN is available at `http://bio.math.berkeley.edu/mjoin/`.

Distance Matrix

A **distance matrix** for a tree T is a matrix D whose entry D_{ij} stands for the mutation distance between i and j.



Distance Matrix

	1	2	3	4	5	6
1	0	6	8	9	12	11
2	6	0	6	7	10	9
3	8	6	0	3	6	5
4	9	7	3	0	5	4
5	12	10	6	5	0	5
6	11	9	5	4	5	0

Table 1: Distance matrix D for the example.

Definitions

Def. A distance matrix D is a **metric** iff D satisfies:

- Symmetric: $D_{ij} = D_{ji}$ and $D_{ii} = 0$.

- Triangle Inequality: $D_{ik} + D_{jk} \geq D_{ij}$.

Def. D is an **additive metric** iff there exists a tree T s.t.

- Every edge has a positive weight and every leaf is labeled by a distinct species in the given set.

- For every pair of i, j, $D_{ij} =$ the sum of the edge weights along the path from i to j.

Also we call such T an **additive tree**.

Neighbor Joining method

Def. We call a pair of two distinct leaves $\{i, j\}$ a **cherry** if there is exactly one intermediate node on the unique path between i and j.

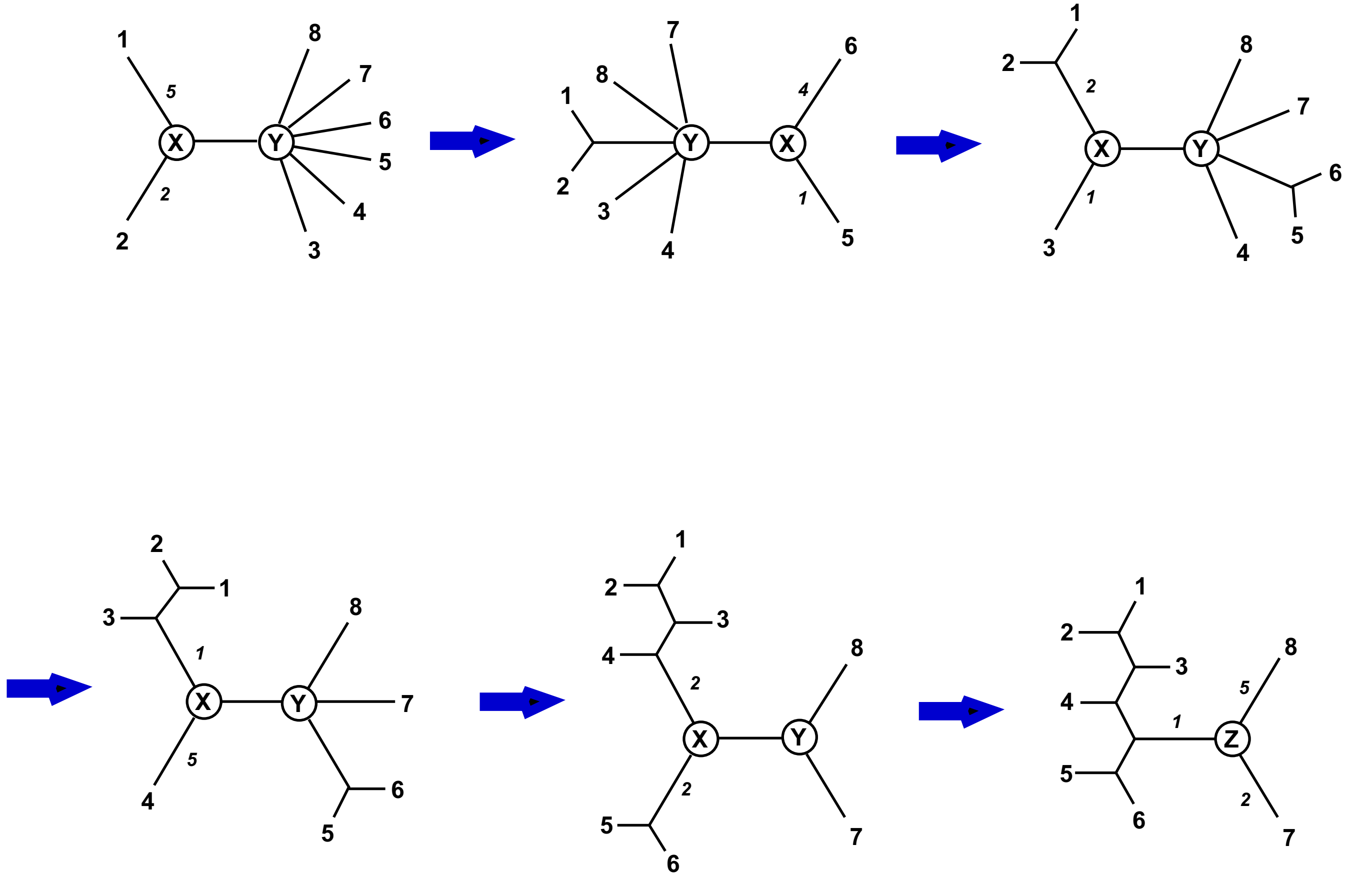
Thm. [Saitou-Nei, 1987 and Studier-Keppler, 1988]

Let $A \in \mathbb{R}^{n \times n}$ such that $A_{ij} = D(ij) - (r_i + r_j)/(n-2)$, where $r_i := \sum_{k=1}^{n} D(ik)$. $\{i^*, j^*\}$ is a cherry in T if $A_{i^* j^*}$ is a minimum for all i and j.

Neighbor Joining Method:

Input. A tree matric D. **Output.** An additive tree T.
Idea. Initialize a star-like tree. Then find a cherry $\{i, j\}$ and compute branch length from the interior node x to i and from x to j. Repeat this process recursively until we find all cherries.

Neighbor Joining Method



The GNJ method

- Extended the Neighbor Joining method with the total branch length of m-leaf subtrees.

- Increasing $2 \leq m \leq n-2$, since there are more data, a reconstructed tree from GNJ method gets closer to the true tree than the Saito-Nei NJ method.

- The time complexity of GNJ method is $O(n^m)$.

Note: If $m = 2$, then GNJ method is the Neighbor Joining method with pairwise distances.

Notation and definitions

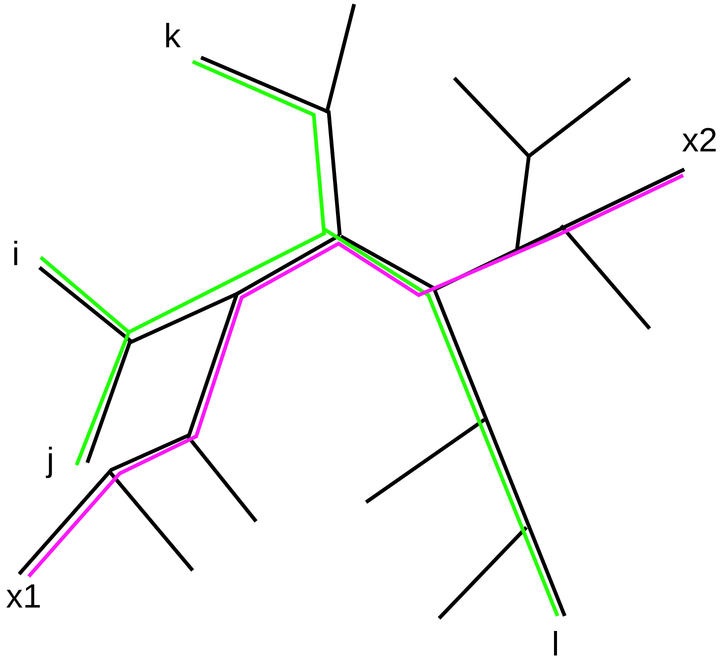
Notation. Let $[n]$ denote the set $\{1, 2, ..., n\}$ and $\binom{[n]}{m}$ denote the set of all m-element subsets of $[n]$.

Def. A **m-dissimilarity map** is a function $D : \binom{[n]}{m} \to \mathbb{R}_{\geq 0}$.

In the context of phylogenetic trees, the map $D(i_1, i_2, ..., i_m)$ measures the weight of a subtree that spans the leaves $i_1, i_2, ..., i_m$.

Denote $D(i_1 i_2 \ldots i_m) := D(i_1, i_2, ..., i_m)$.

Weights of Subtrees in T



$D(ijkl)$ is the total branch length of the subtree in green. Also $D(x_1x_2)$ is the total branch length of the subtree in pink and it is also a pairwise distance between x_1 and x_2.

Thm. [Levy, Y., Pachter, 2005] Let D_m be an m-dissimilarity map on n leaves of a tree T, $D_m : \binom{[n]}{m} \to \mathbb{R}_{\geq 0}$ corresponding m-subtree weights, and define

$$\mathbf{S(ij) := \sum_{X \in \binom{[n] \setminus \{i,j\}}{m-2}} D_m(ijX).}$$

Then $S(ij)$ is a tree metric.

Furthermore, if T' is based on this tree metric $S(ij)$ then T' and T have the same tree topology and there is an invertible linear map between their edge weights.

Note. This means that if we reconstruct T', then we can reconstruct T.

Neighbor Joining with Subtree Weights

Input: n DNA sequences and an integer $2 \leq m \leq n-2$.

Output: A phylogenetic tree T with n leaves.

1. Compute all m-subtree weights via the ML method.

2. Compute $S(ij)$ for each pair of leaves i and j.

3. Apply Neighbor Joining method with a tree metric $S(ij)$ and obtain additive tree T'.

4. Using a one-to-one linear transformation, obtain a weight of each internal edge of T and a weight of each leaf edge of T.

Complexity

Lemma. [Levy, Pachter, Y.] If $m \geq 3$, the time complexity of this algorithm is $O(n^m)$, where n is the number of leaves of T and if $m = 2$, then the time complexity of this algorithm is $O(n^3)$.

Sketch of Proof: If $m \geq 3$, the computation of $S(ij)$ is $O(n^m)$ (both steps are trivially parallelizable). The subsequent neighbor-joining is $O(n^3)$ and edge weight reconstruction is $O(n^2)$. If $m = 2$, then the subsequent neighbor-joining is $O(n^3)$ which is greater than computing $S(ij)$. So, the time complexity is $O(n^3)$.

Note: The running time complexity of the algorithm is $O(n^3)$ for both $m = 2$ and $m = 3$.

Cherry Picking Theorem

Thm. [Levy, Pachter, Y.] Let T be a tree with n leaves and no nodes of degree 2 and let m be an integer satisfying $2 \leq m \leq n-2$. Let $D : \binom{[n]}{m} \to \mathbb{R}_{\geq 0}$ be the m-dissimilarity map corresponding to the weights of the subtrees of size m in T. If $Q_D(a^*b^*)$ is a minimal element of the matrix

$$Q_D(ab) = \left(\frac{n-2}{m-1}\right) \sum_{X \in \binom{[n-i-j]}{m-2}} D(ijX) - \sum_{X \in \binom{[n-i]}{m-1}} D(iX) - \sum_{X \in \binom{[n-j]}{m-1}} D(jX)$$

then $\{a^*, b^*\}$ is a cherry in the tree T.

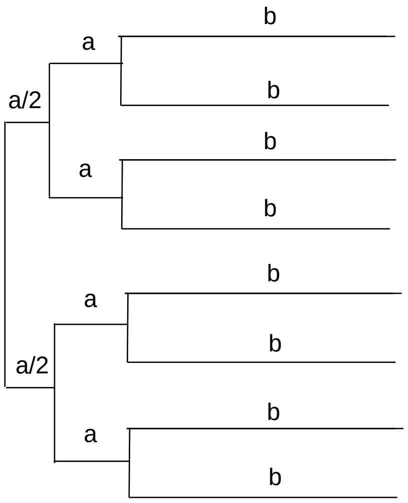
Note. The theorem by Saitou-Nei and Studier-Keppler is a corollary from Cherry Picking Theorem.

Simulation Results

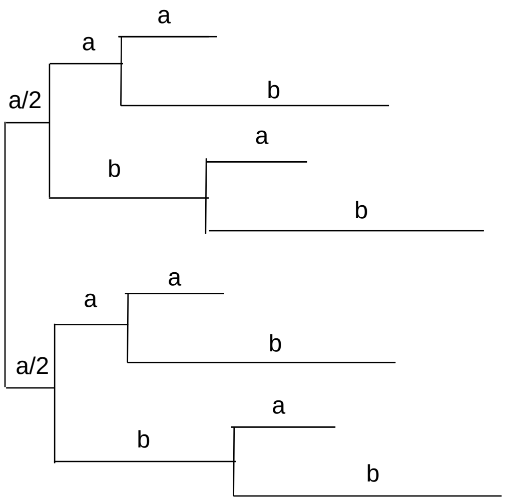
With the Juke Cantor model.

Consider two tree models...

Modeled from Strimmer and von Haeseler.



T1



T2

We generate 500 replications with the Jukes-Cantor model via a software `evolver` from `PAML` package.

The number represents a percentage which we got the same tree topology.

l	a/b	m=2	m=3	m=4	fastDNAml
500	0.01/0.07	68.2	76.8	80.4	74.8
	0.02/0.19	54.2	61.2	73.6	55.6
	0.03/0.42	10.4	12.6	23.8	12.6
1000	0.01/0.07	94.2	96	97.4	96.6
	0.02/0.19	87.6	88.6	96.2	88
	0.03/0.42	33.4	35	52.4	33.6

Table 2: Success Rates for the model T_1.

l	a/b	m=2	m=3	m=4	fastDNAml
500	0.01/0.07	84.4	86	85.6	88.4
	0.02/0.19	68.2	72	73.2	88.4
	0.03/0.42	18.2	29.2	36.2	87.4
1000	0.01/0.07	95.6	97.8	97.4	99.4
	0.02/0.19	88.4	89.6	93.4	99.8
	0.03/0.42	40	48.2	57.6	96.6

Table 3: Success Rates for the model T_2.

Applications of GNJ method

The EMGNJ algorithm

The GNJ method: in 2005, Levy, Y., and Pachter introduced the **generalized neighbor-joining (GNJ) method**, which reconstructs a phylogenetic tree based on comparisons of subtrees rather than pairwise distances

The EMGNJ algorithm (*the Algebraic Biology*, 2005): iterates between the EM algorithm for estimating substitution rates and the generalized NJ method for phylogenetic tree reconstruction.

Simulation Results

We implemented subroutines of the EMGNJ algorithm with $m = 4$ under the JC model.

S-locus receptor kinase (SRK)

In pollen, Plant self-incompatibility (SI) specificity is determined by the S-locus cysteine-rich protein gene (SCR), which encodes small secreted hydrophilic and positively charged proteins of 50 to 59 amino acids.

Both SRK and SCR are members of large families of genes that are expressed in a variety of plant tissues.

Maturation of the flower in self-incompatible crucifers is accompanied by the insertion of SRK into the plasma membrane of stigma epidermal cells and of SCR into the pollen coat.

“Recognition and rejection of self in plant reproduction” by JB Nasrallah *Science*, **296**, (2002) p 305 – 308.

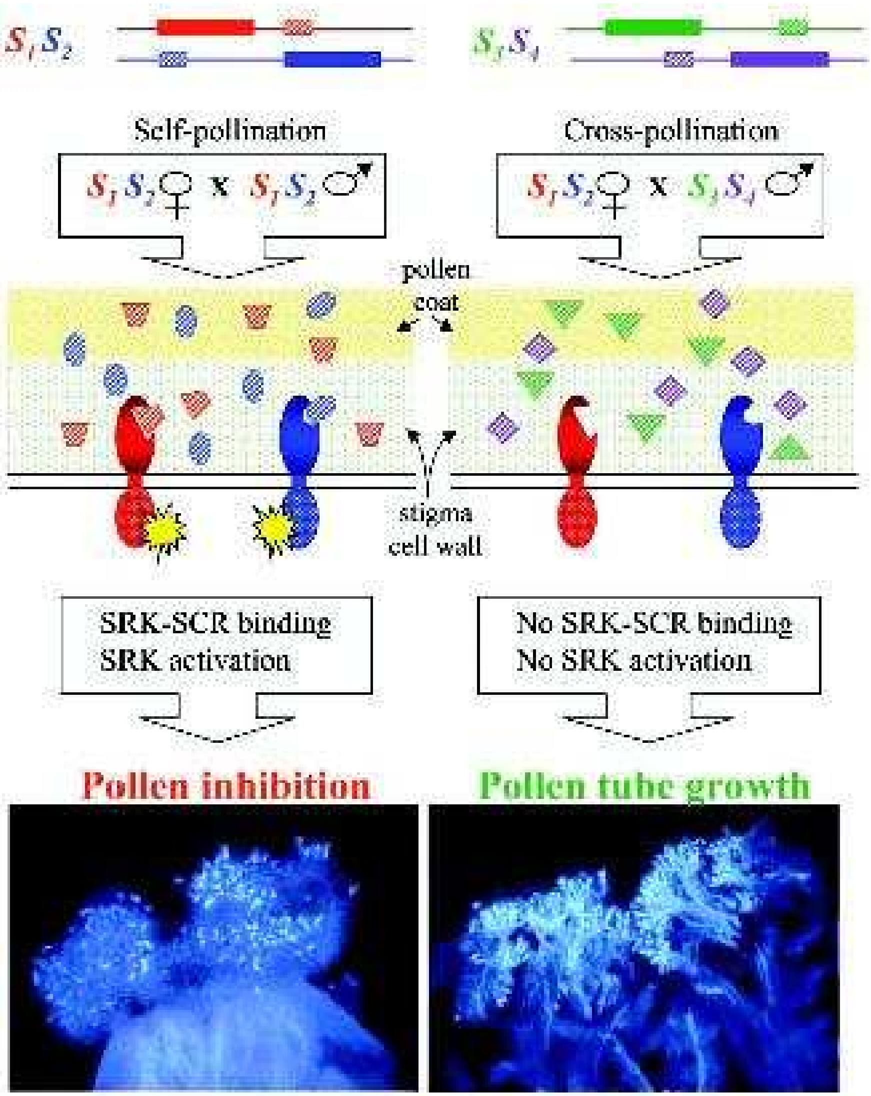


Figure 1: Nasrallah (2002), *Nature*

Find the phylogenetic tree for 21 different species' *S-locus* receptor kinase (SRK) sequences involved in the self/nonself discriminating self-incompatibility system of the mustard family (Sainudiin et al, 2005).

Symmetric difference ($\mathbf{\Delta}$) between $10,000$ trees sampled from the likelihood function via MCMC and the trees reconstructed by 5 methods.

DNAml(A) is a basic search with no global rearrangements, whereas DNAml(B) applies a broader search with global rearrangements and randomize input order of sequences 100 times.

A = sub-routine of the EMGNJ method, B = Saitou-Nei NJ method, C = fastDNAml, D = DNAml(A), F = DNAml(B), and G = TrExML.

$\boldsymbol{\Delta}$	A	B	C	D	F	G
0	0	0	0	2	3608	0
2	77	0	0	1	471	0
4	3616	171	6	3619	5614	0
6	680	5687	5	463	294	5
8	5615	4134	3987	5636	13	71
10	12	8	5720	269	0	3634
12	0	0	272	10	0	652
14	0	0	10	0	0	5631
16	0	0	0	0	0	7

The result tree via the EMGNJ method is much better than the Saito-Nei NJ metho dTrExML and fastDNAml.

Stochastic NJ Importance Sampling method

The Moore-rejection samplers are a class of rejection samplers that can be applied to target a density over a compact domain with a well-defined interval extension. Interval arithmetics are binary operations over intervals instead of using real numbers.

Using the Moore-rejection samplers via interval methods we can sample trees via the GNJ method. We call this method, the **Bayesian interval generalized neighbor-joining method (BIGNJ)**.

The outline of this method is the following.

1. Take samples for each subtrees with size m via the samplers.

2. Consider the set of samples as an estimation of a probability distribution for each subtree.

3. Instead of taking a total branch length of each subtree, take the estimation of a probability distribution for each subtree weight.

4. Run the GNJ method with the set of estimations of probability distributions.

- D. Levy (Math, Berkeley), L. Pachter (Math, Berkeley), and R. Yoshida, "Beyond Pairwise Distances: Neighbor Joining with Phylogenetic Diversity Estimates" the Molecular Biology and Evolution, Advanced Access, November 9, (2005).

- A. Hobolth (Bioinformatics, NCSU) and R. Yoshida, "Maximum likelihood estimation of phylogenetic tree and substitution rates via generalized neighbor-joining and the EM algorithm", *Algebraic Biology 2005, Computer Algebra in Biology*, edited by H. Anai and K. Horimoto, vol. 1 (2005) p41 - 50, Universal Academy Press, INC.. (Also available at arXiv:q-bio.QM/0511034.)

- R. Sainudiin (Statistics, Oxford) and R. Yoshida, "Applications of Interval Methods to Phylogenetic trees" *Algebraic Statistics for Computational Biology* edited by Lior Pachter and Bernd Sturmfels, (2005) Cambridge University Press, p359 - 374.

Thank you....