Maximum likelihood estimation of phylogenetic trees and substitution rates

Ruriko Yoshida
Dept. of Mathematics Duke University

Joint work with Asger Hobolth

www.math.duke.edu/~ruriko

January 20th, 2006

Challenge

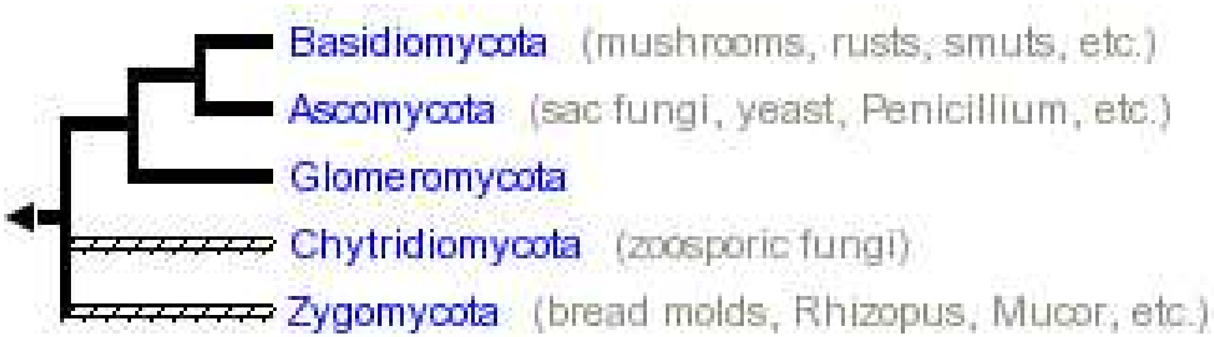
We would like to assemble the fungi tree of life.

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....

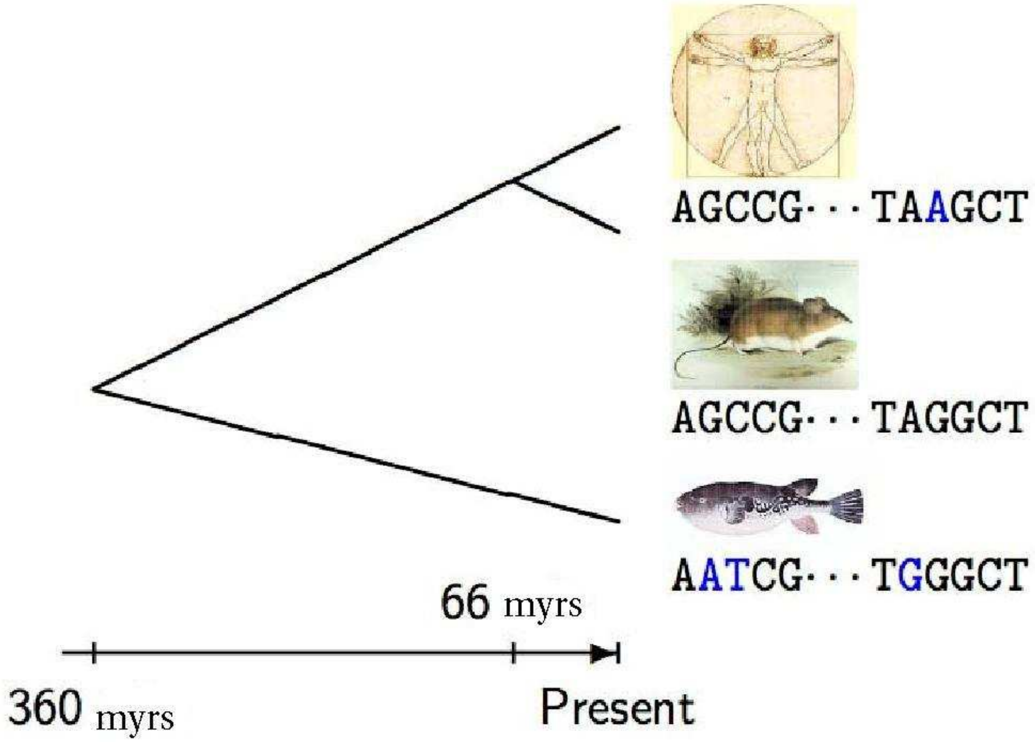

Basidiomycota (mushrooms, rusts, smuts, etc.)
Ascomycota (sac fungi, yeast, Penicillium, etc.)
Glomeromycota
Chytridiomycota (zoosporic fungi)
Zygomycota (bread molds, Rhizopus, Mucor, etc.)

`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 to estimate substitution rate and 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 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

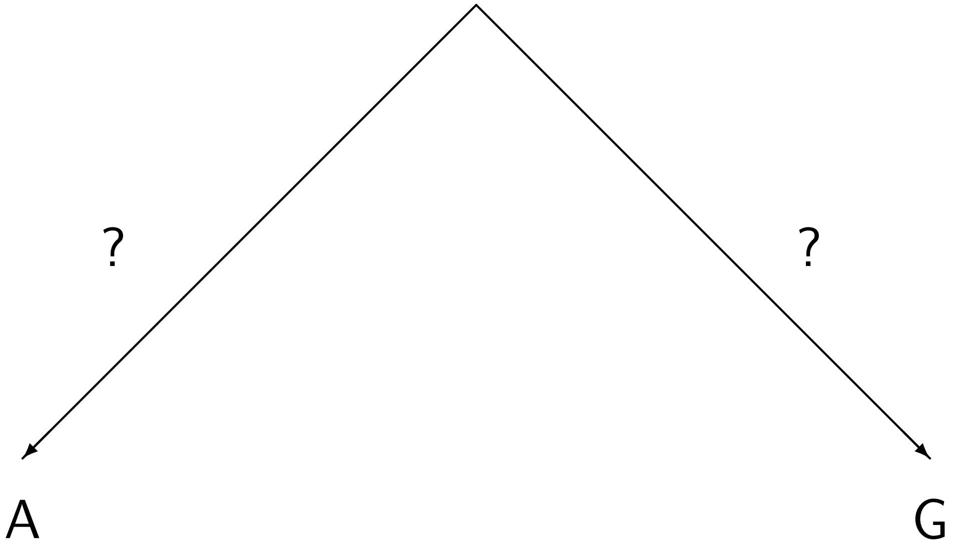
- The GNJ method uses more sequence information: the resulting tree should be more accurate than the NJ method.
- The computational time: polynomial in terms of the number of DNA sequences.

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

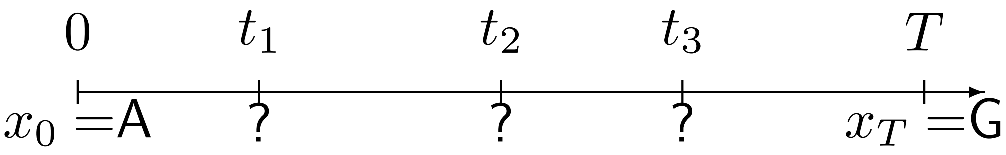
The EM algorithm for estimating substitution rates

Pairwise sequences

Suppose we have a pair of sequences at a single site such that:



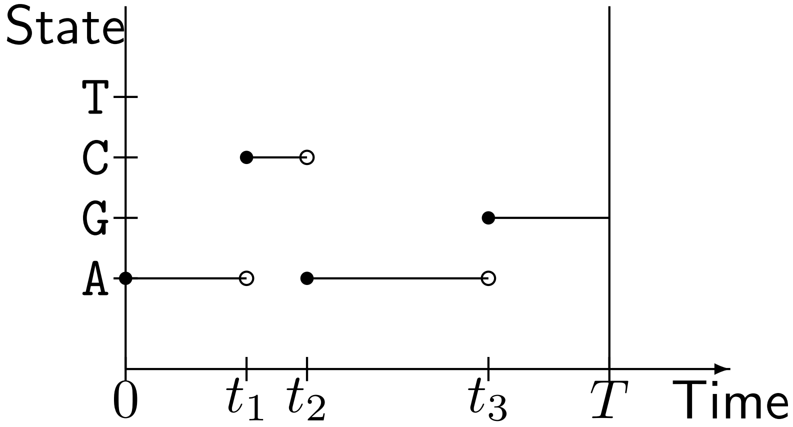
Assuming time reversibility....



Complete observation

Suppose we have the **complete observation** of continuous time Markov chain $x = \{x_t : 0 \leq t \leq T\}$ on the state space $\Sigma = \{A, C, G, T\}$.

Example:



We wish to estimate the substitution rate matrix Q using Maximum Likelihood.

Let $Q = (Q(a,b))_{a,b\in\Sigma} \in \mathbb{R}^{4\times 4}$.

The waiting time in a state a has the density $-Q(a,a)\exp(-Q(a,a)t)$ and Prob$(a \to b) = -Q(a,b)/Q(a,a)$.

Thus we can write the likelihood function of the example as:

$$e^{Q(A,A)(t_1+(t_3-t_2))+Q(G,G)(T-t_3)+Q(C,C)(t_2-t_1)}Q(A,G)Q(A,C)Q(C,A).$$

So, we can write the likelihood in terms of $T(a)$, the total time spent in state a, and $N(a,b)$, the number of substitutions of state a with b.

More generally...

In general, if Q is parametrized by θ, $Q = Q_\theta$, and with $x = \{x(t) : 0 \leq t \leq T\}$, the MLE problem for a complete observation is:

$$\max \quad \mathbf{L}(\theta; \mathbf{x}) \; = \; \left[\prod_{\mathbf{a} \in \boldsymbol{\Sigma}} \prod_{\mathbf{b} \neq \mathbf{a}} \mathbf{Q}_\theta(\mathbf{a}, \mathbf{b})^{\mathbf{N}(\mathbf{a}, \mathbf{b})} \right] \left[\prod_{\mathbf{a} \in \boldsymbol{\Sigma}} \exp(\mathbf{Q}_\theta(\mathbf{a}, \mathbf{a}))^{\mathbf{T}(\mathbf{a})} \right]$$

such that $\theta \in \Theta$.

The log-likelihood for a complete observation becomes

$$\log \mathbf{L}(\theta; \mathbf{x}) = \sum_{\mathbf{a} \in \boldsymbol{\Sigma}} \sum_{\mathbf{b} \neq \mathbf{a}} \mathbf{N}(\mathbf{a}, \mathbf{b}) \log \mathbf{Q}_\theta(\mathbf{a}, \mathbf{b}) + \sum_{\mathbf{a} \in \boldsymbol{\Sigma}} \mathbf{T}(\mathbf{a}) \mathbf{Q}_\theta(\mathbf{a}, \mathbf{a}).$$

The GTR model

Consider the general time reversible (GTR) model.

Let $\pi_a, \ a \in \Sigma, \ \sum_a \pi_a = 1$, be the stationary distribution of the Markov chain.

The GTR model has substitution rate matrix:

$$Q_\theta = \begin{bmatrix} \cdot & \theta_{AG}\pi_G & \theta_{AC}\pi_C & \theta_{AT}\pi_T \\ \theta_{AG}\pi_A & \cdot & \theta_{GC}\pi_C & \theta_{GT}\pi_T \\ \theta_{AC}\pi_A & \theta_{GC}\pi_G & \cdot & \theta_{CT}\pi_T \\ \theta_{AT}\pi_A & \theta_{GT}\pi_G & \theta_{CT}\pi_C & \cdot \end{bmatrix}$$

where the diagonal elements are such that each row sums to zero.

The 6 unknown parameters are $\theta = (\theta_{AG}, \theta_{AC}, \theta_{AT}, \theta_{GC}, \theta_{GT}, \theta_{CT})$.

Missing data problem

Problem: If we only observe $x(0)$ and $x(T)$?

Note: The complete log-likelihood is maximized for

$$\theta^*_{\mathbf{ab}} = \frac{\mathbf{N(a, b) + N(b, a)}}{\pi_{\mathbf{b}}\mathbf{T(a)} + \pi_{\mathbf{a}}\mathbf{T(b)}}, \quad \mathbf{a \neq b.} \tag{1}$$

The EM algorithm:

1. (Expectation Step) Calculate $T(a)^* := E[T(a) : x(0), x(T)]$ and $N(a, b)^* := E[N(a, b) : x(0), x(T)]$.

2. (Maximization Step) Substitute $T(a)^*$ and $N(a, b)^*$ into Equation (1).

Iterate between Step 1 and Step 2 until convergence.

[due to the Chapman-Kolmogorov equation, Hobolth and Jensen, 2005]

Denote the transition matrix $P(t) = \exp(Qt)$.

- Time spent in state a

$$\mathrm{E}[\mathbf{T}(\mathbf{a})|\mathbf{x}(\mathbf{0}) = \mathbf{i}, \mathbf{x}(\mathbf{T}) = \mathbf{j}] = \int_{\mathbf{0}}^{\mathbf{T}} \mathbf{P}_{\mathbf{ia}}(\mathbf{t})\mathbf{P}_{\mathbf{aj}}(\mathbf{T} - \mathbf{t})\mathbf{dt}/\mathbf{P}_{\mathbf{ij}}(\mathbf{T}).$$

- Number of transitions between states a and b

$$\mathrm{E}[\mathbf{N}(\mathbf{a}, \mathbf{b})|\mathbf{x}(\mathbf{0}) = \mathbf{i}, \mathbf{x}(\mathbf{T}) = \mathbf{j}] = \mathbf{Q}(\mathbf{a}, \mathbf{b}) \int_{\mathbf{0}}^{\mathbf{T}} \mathbf{P}_{\mathbf{ia}}(\mathbf{t})\mathbf{P}_{\mathbf{bj}}(\mathbf{T} - \mathbf{t})\mathbf{dt}/\mathbf{P}_{\mathbf{ij}}(\mathbf{T}).$$

Multiple sequences

Assuming that the multiple alignment is given. We fix the tree topology for a tree T with n leaves. Note that there are $2n-3$ edges in T.

The single site complete log-likelihood becomes

$$\log \mathbf{L}(\theta; \mathbf{x}) = \sum_{\mathbf{i=1}}^{\mathbf{2n-3}} \Big(\sum_{\mathbf{a}\in\boldsymbol{\Sigma}} \sum_{\mathbf{b}\neq\mathbf{a}} \mathbf{N^i(a,b)} \log \mathbf{Q^i(a,b)} + \sum_{\mathbf{a}\in\boldsymbol{\Sigma}} \mathbf{T^i(a)Q^i(a,a)} \Big)$$

where $T^i(a)$ is the total time spent in state a on edge i and $N^i(a,b)$ is the number of transitions from a to b on edge i.

Apply the previous equations and Felsenstein's pruning algorithm to solve the problem.

The GNJ method

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

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.

Let $D(ij)$ be a pairwise distance between i and j.

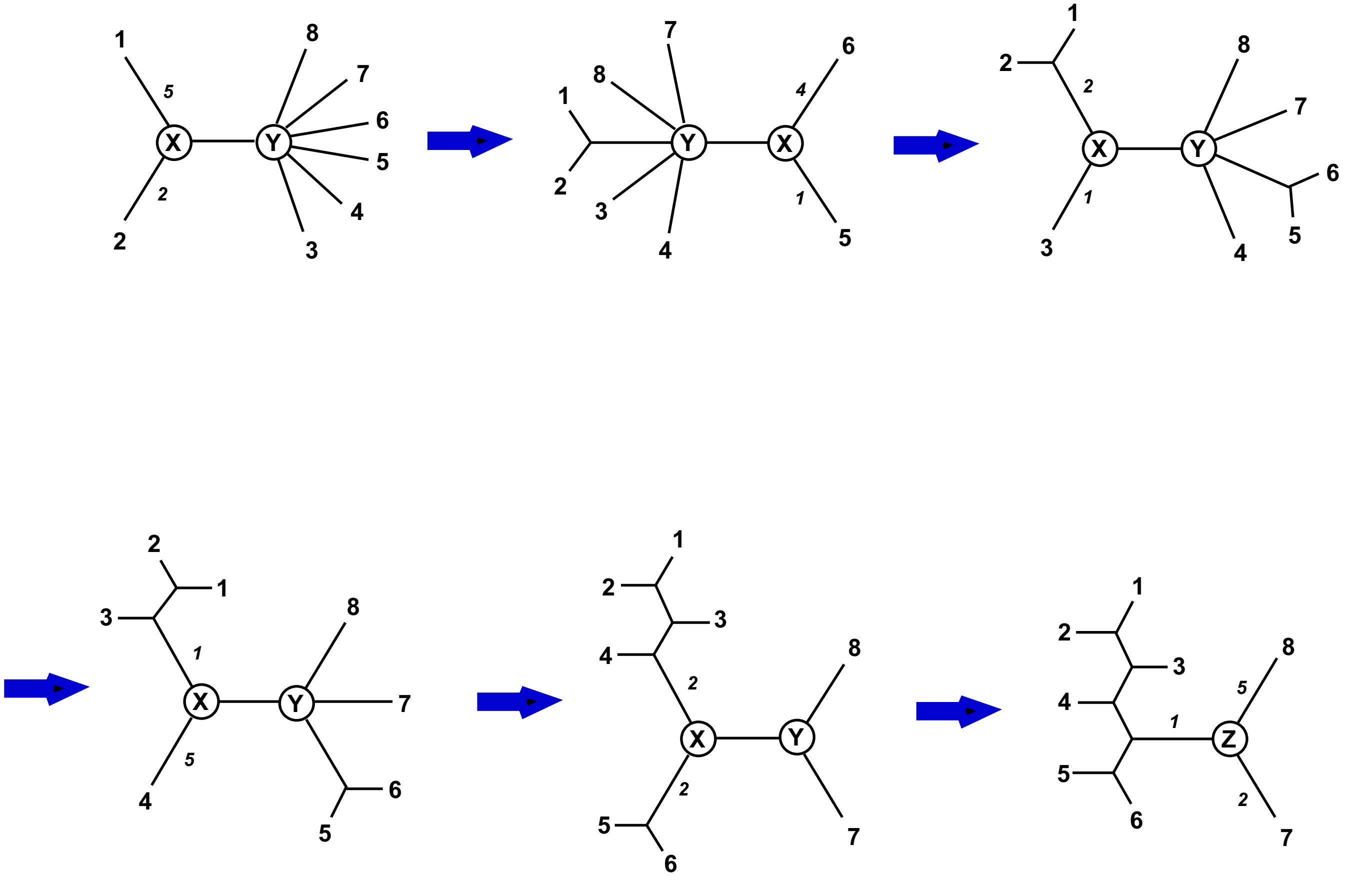
Thm. [Saitou-Nei, 1987 and Studier-Keppler, 1988]

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 over all i and j.

Neighbor Joining Method:

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

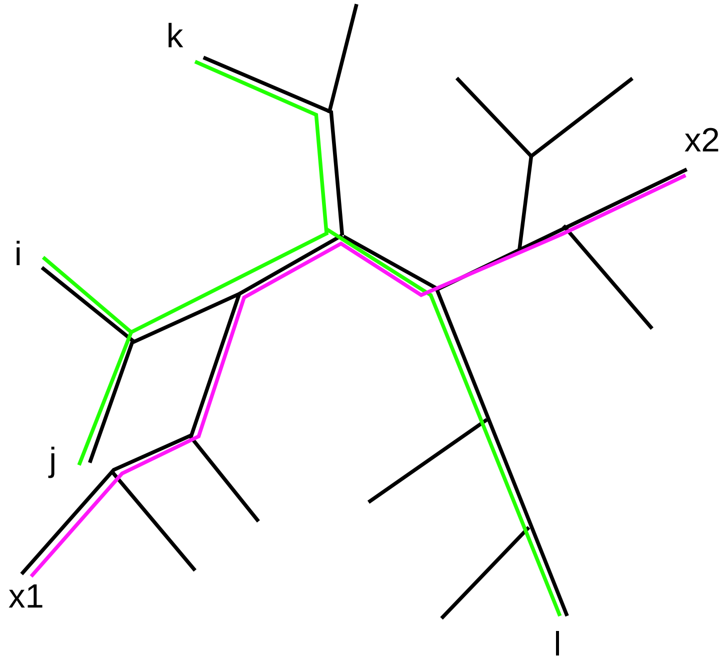
Notation. Let $[n]$ denote the set $\{1, 2, ..., n\}$ and $\binom{[n]}{m}$ denote the set of all m-element subsets of $[n]$.

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

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, $D_m : \binom{[n]}{m} \rightarrow \mathbb{R}_{\geq 0}$ 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.

The EMGNJ Algorithm

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

Output: The GTR rates and a phylogenetic tree.

1. Estimate stationary distribution from empirical frequencies.

2. Reconstruct tree using the GNJ method under the JC69 model.

3. Estimate GTR substitution rates and edge lengths from current tree via the EM algorithm.

4. Reconstruct tree using the GNJ method and current GTR rates.

5. If likelihood is not improved return current tree and GTR rates; otherwise go to 3.

Simulation Results

We implemented subroutines of the EMGNJ algorithm, Step 3 and Step 4 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.

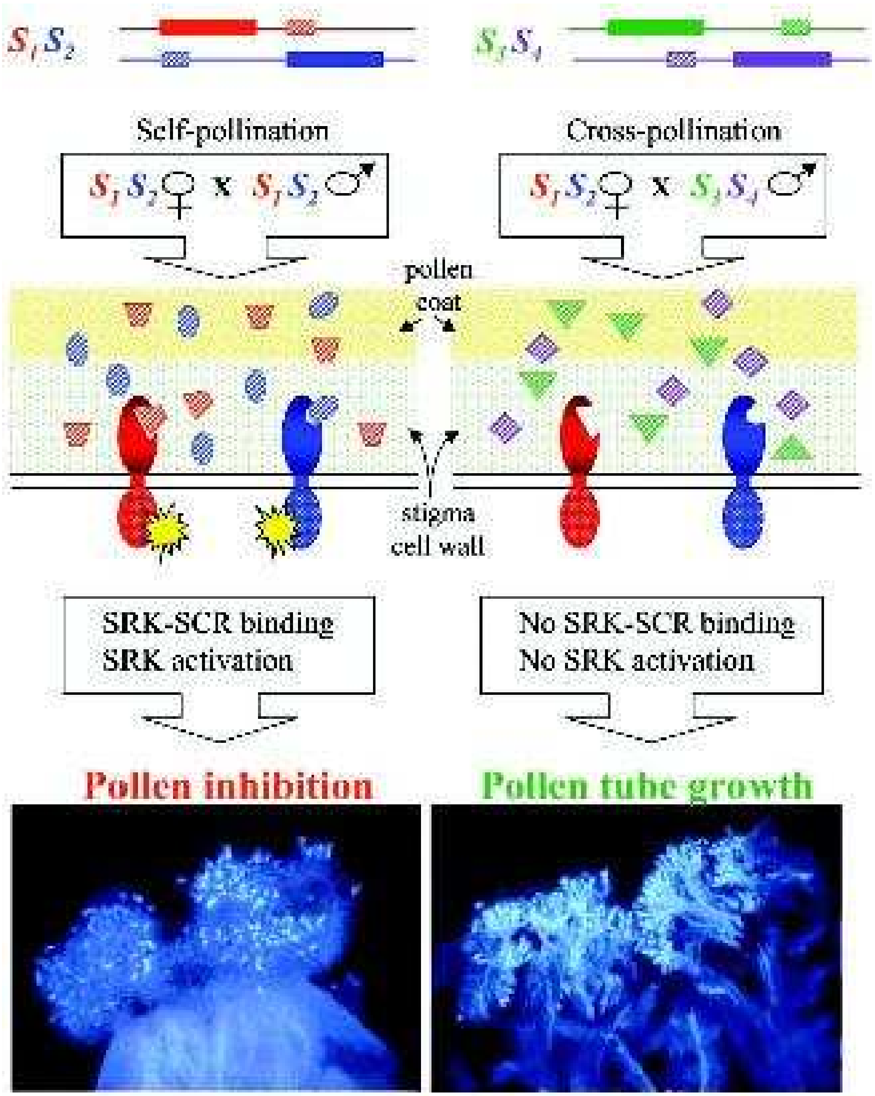


Figure 1: Nasrallah (2002), *Nature*

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.

$\mathbf{\Delta}$	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.

A unifying framework: Algebraic Statistics

What is Algebraic Statistics?

Algebraic Statistics is to apply computational commutative algebraic techniques to statistical problems.

The algebraic view of discrete statistical models has been applied in many statistical problems, including:

- conditional inference [Diaconis and Sturmfels 1998]
- disclosure limitation [Sullivant 2005]
- the maximum likelihood estimation [Hosten et al 2004]
- parametric inference [Pachter and Sturmfels 2004]
- phylogenetic invariants [Allman and Rhodes 2003, Eriksson 2005, etc].

Algebraic statistical models

An **algebraic statistical model** arises as the image of a polynomial map

$$\mathbf{f} \ : \ \mathbb{R}^d \rightarrow \mathbb{R}^m \ , \quad \theta = (\theta_1, \ldots, \theta_d) \ \mapsto \ \big(p_1(\theta), p_2(\theta), \ldots, p_m(\theta)\big).$$

The unknowns $\theta_1, \ldots, \theta_d$ represent the model parameters.

In the view of algebraic geometry, statistical models are **algebraic varieties**, sets of points where all given polynomials vanish at the same time.

Note: The phylogenetic models are also algebraic varieties.

Note: The MLE problem is a polynomial optimization problem over the image of $\mathbf{f}$.

Pairwise sequences with a site

Suppose we have the observed data and assume that we know all information about t_i. We want to estimate a state in each t_i.

Recall: The transition probability from a to b is $-Q_\theta(a,b)/Q_\theta(a,a)$.

From the view of algebra, the MLE for an observation is the following: Let $\sigma = |\Sigma| = 4$. Consider an algebraic statistic model such that:

$$\mathbf{f} : \mathbb{R}^{\binom{\sigma}{2}} \longrightarrow \mathbb{R}^{\sigma(\sigma-1)}, \ \theta \mapsto \big(p_{1,2}(\theta), \ldots, p_{\sigma,\sigma-1}(\theta)\big),$$

where $p_{a,b}(\theta) = \text{prob}(a \rightarrow b) = -Q_\theta(a,b)/(\sigma Q_\theta(a,a))$.

The likelihood for a complete observation becomes

$$\mathbf{L(\theta; x)} = \left[\prod_{\mathbf{a \in \Sigma}} \prod_{\mathbf{b \neq a}} \mathbf{p_{a,b}^{N(a,b)}}\right].$$

Note: The image of $\mathbf{f}$ is a hyper-surface over $\mathbb{R}^{\sigma(\sigma-1)}$.

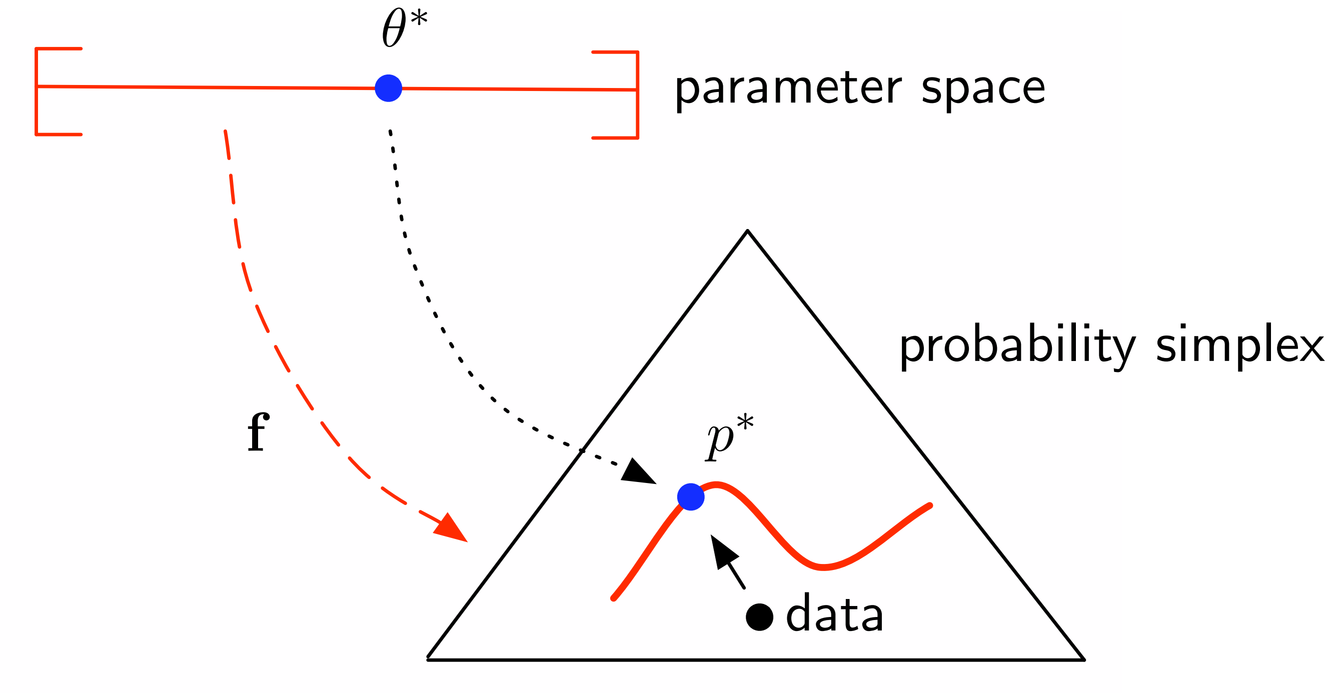


Figure 2: θ^* is the global maxima and p^* is an image under $\mathbf{f}$.

Likelihood variety

Definition: The **likelihood variety** V of the model $\mathbf{f}$ with respect to the data is the closure of the set of critical points of $\ell = \log L$ in $\mathbb{R}^{\binom{\sigma}{2}}$.

Every local and global maximum of the complete log-likelihood is a solution of the **critical equations**

$$\frac{\partial \ell}{\partial \theta_{AG}} = \frac{\partial \ell}{\partial \theta_{AC}} = \cdots = \frac{\partial \ell}{\partial \theta_{CT}} = 0.$$

Thus the likelihood variety is the closure of the set of points which all $\frac{\partial \ell}{\partial \theta_{ab}}$ vanish. By clearing the denominators and using **Gröbner bases** we can compute the likelihood variety.

Note: The likelihood variety of the complete observation contains the global maxima θ^*.

However, we do not have a complete observation. Thus we use the EM algorithm to estimate a point close to the likelihood variety.

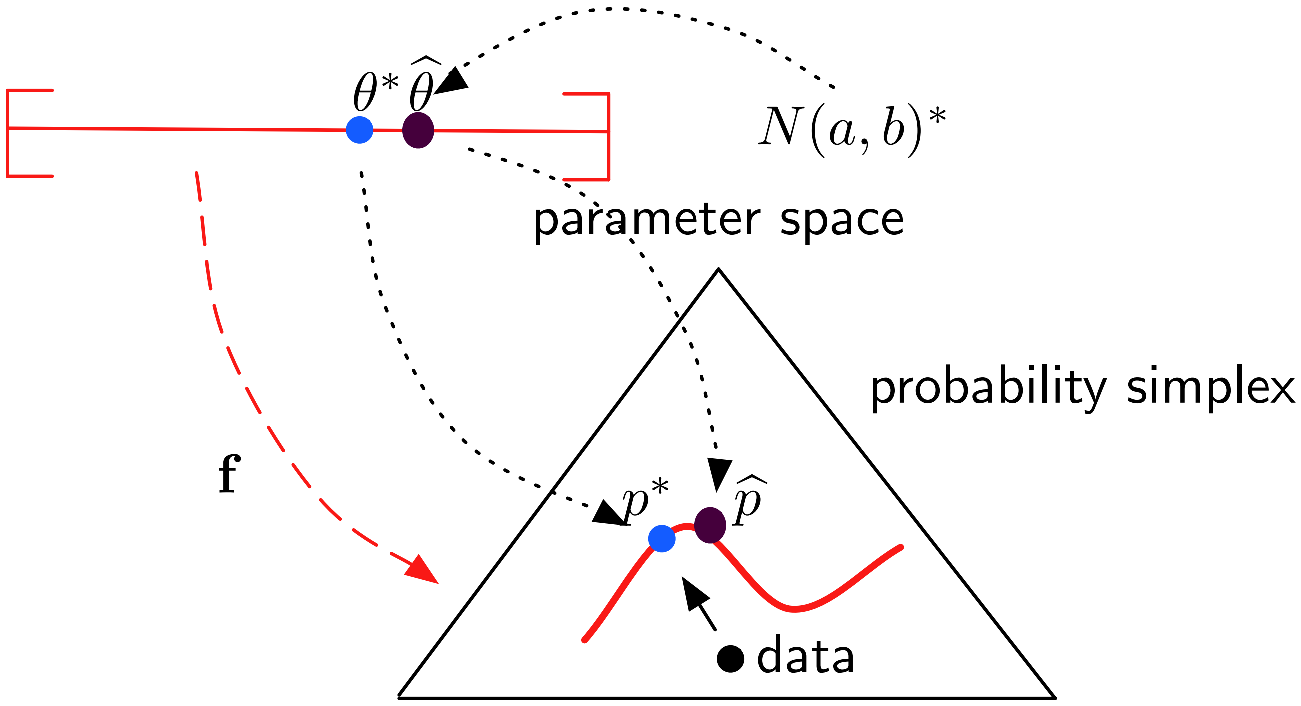


Figure 3: $\widehat{\theta}$ is estimated by E-step and $\widehat{p}$ is an image under $\mathbf{f}$.

Commutative Algebra Applications in Phylogenetics

Algebraic varieties are well studied in algebraic geometry.

Any algebraic variety can be described by the set of all polynomials which vanish at all points in the variety, which is called an **ideal** I over a polynomial ring. (I is an infinite set. However, I can be described by a finite set of polynomials, a **basis**, such as a **Gröbner basis**.)

Using a Gröbner basis for an ideal, one can apply to the MLEs problems (e.g. Hosten et al 2004).

One can describe a hypersurface of an algebraic statistics model using invariants (e.g. Allman and Rhodes 2003, Eriksson 2005).

One can find more applications of algebra to computational biology at our new book **Algebraic Statistics for Computational Biology** edited by Pachter and Sturmfels, Cambridge University Press 2005.

- 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....