

# Phylogenetics tutorial



Elizabeth S. Allman  
University of Alaska  
Fairbanks 

---

September 14, 2008  
SAMSI  
Opening workshop

# Outline

1. Overview of phylogenetic methods
  - ▶ Maximum parsimony
  - ▶ Distance methods
  - ▶ Model-based methods (ML and Bayesian framework)
2. Models of site substitution
3. Identifiability of parameters
  - ▶ background including discussion of information content of  $k$ -marginals
  - ▶ Numerous theorems

Much of the mathematical work discussed today is joint work with



John Rhodes

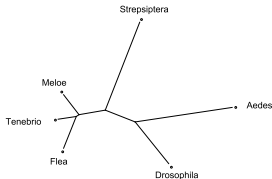
# Inference Problem:

Given aligned biological sequences, presumed to have arisen from a common ancestral sequence, infer their evolutionary history.

<b>Strepsiptera</b>	AAGCTCATTAAATCGCTTTGGTTCCTTAGATAGTTGGAT...
<b>Aedes</b>	AGGCTCAGTATAACACTATAATTTACAAGATCATTGGAT...
<b>Drosophila</b>	AGGCTCATTATATCATTATGGTTCCTTAGATCGTTGGAT...
<b>Flea</b>	TGGCTCATTATATCATTATGGTTCATTAGATCGTTGGAT...
<b>Meloe</b>	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT...
<b>Tenebrio</b>	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT...



H=0.01



## Mathematical aspects of Phylogenetics:

- Modeling
- Combinatorics
- Algebra
- ??
- Statistics
- Algorithms
- Geometry

For phylogenetic inference,  
the data are *observed pattern frequencies* in aligned sequences:

Strepsiptera	AAGCTCATTAAATCGCTTTGGTTCCTTAGATAGTTGGAT...
Aedes	AGGCTCAGTATAACACTATAATTTACAAGATCATTGGAT...
Drosophila	AGGCTCATTATATCATTATGGTTCCTTAGATCGTTGGAT...
Flea	TGGCTCATTATATCATTATGGTTCATTAGATCGTTGGAT...
Meloe	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT...
Tenebrio	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT...

$$\hat{p}_{AAAAAA} = \frac{\# \text{ observations of } AAAAAA}{\text{sequence length}}, \text{ etc.}$$

which, assuming a model of molecular evolution along a tree, are  
estimators for the true *joint distribution*  $p_{AAAAAA}$ , etc.

Major approaches in current use:

- Maximum Parsimony
- Distance Methods (NJ)
- model-based methods {
  - Maximum Likelihood (ML)
  - Bayesian Methods

# Maximum Parsimony

“The best tree (or trees) is the one that requires the fewest number of mutations.”

The minimal number of mutations for a particular tree is known as its *parsimony score*.

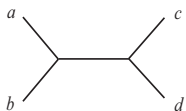
Historically, MP was used to construct trees from phenotypic data (beak size, webbed feet, etc.)

## Example:

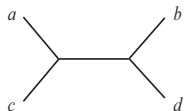
- a: ATTAGGTACATGATTAG
- b: ATTCGGTACATGATTAG
- c: ATTCGCTACATGATCCG
- d: ATTTGCTACATGTTCCG

(recall: data is  $\hat{p}_{AAAA} = 3/17$ ,  $\hat{p}_{ACCT} = 1/17$ , ...)

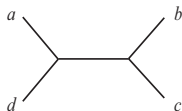
Which of three possible trees best relates the taxa *a*, *b*, *c*, and *d*?



$T_1$



$T_2$



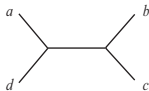
$T_3$

If mutation is rare

— best tree minimizes substitutions

- a: ATT**T**GGTACATGATTAG
- b: ATT**C**GGTACATGATTAG
- c: ATT**C**GCTACATGATCCG
- d: ATT**T**GCTACATGTTCCG

Site 4 is evidence for  $T_3$

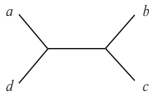


If mutation is rare

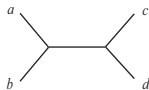
— best tree minimizes substitutions

a: ATTTG**G**TACATGAT**TAG**  
b: ATT**CG****G**TACATGAT**TAG**  
c: ATT**CG****C**TACATGAT**CCG**  
d: ATTT**G****C**TACATGTT**CCG**

Site 4 is evidence for  $T_3$



Sites 6, 15, 16 are evidence for  $T_1$



## Algorithm:

- ▶ Enumerate all unrooted trees relating  $n$  taxa.
- ▶ For each tree, compute parsimony score.
- ▶ Return most parsimonious tree(s).

Number of unrooted  $n$ -taxon trees is  $(2n - 5)!! = 1 \cdot 3 \cdot 5 \cdots (2n - 5)$

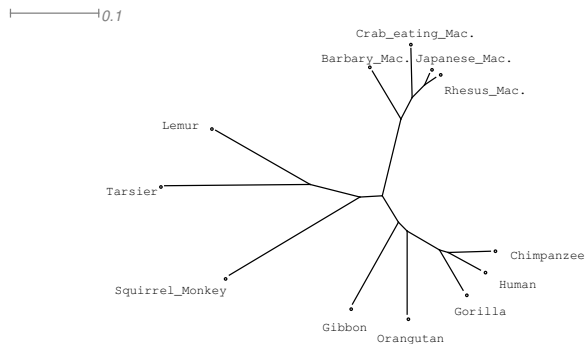
Finding optimal tree is NP-hard.

Heuristic searches are needed.

In practice, MP is performed only on **parsimony-informative** sites, those for which two character states occur at least twice.

# Distance Methods

- used to create unrooted metric trees



Edge length might be interpreted as:

- ▶ measure of difference between sequences (Hamming distances)
- ▶ roughly proportional to time
- ▶ the amount of mutation that occurred along an edge (requires probabilistic model).
- ▶ something else ...

Distance methods begin with a pairwise measure of dissimilarity between species.

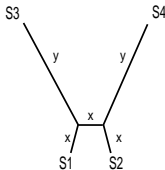
Pairwise distances between taxa are calculated from DNA sequence data in a variety of ways (Hamming distance, probabilistic models, ...)

	Gor	Orangu	Human	Chimp	Gibbon	CEMac	Lemur	BMacaq	JMacaq	...
Gor	0	.1890	.1100	.1130	.2150	.3150	.3470	.2850	.2740	...
Orangu		0	.1790	.1920	.2110	.3170	.3440	.2790	.2890	...
Human			0	.0940	.2050	.2920	.3720	.3040	.2680	...
Chimp				0	.2140	.3240	.3720	.2920	.2850	...
Gibbon					0	.3080	.3540	.2860	.2930	...
CEMac						0	.3610	.1350	.0880	...
Lemur							0	.3430	.3360	...
BMacaq								0	.1370	...
JMacaq									0	...
:										...
:										...

The fast *Neighbor Joining Method* is based on the 4-point condition to determine which two taxa are neighbors.

For example, if  $S1$  and  $S3$  are neighbors, then

$$d(S1, S3) + d(S2, S4) \leq d(S1, S2) + d(S3, S4).$$



$S1$  and  $S3$  are neighbors,  
though  $S1$  and  $S2$  are separated by a shorter distance.

Pairwise distances from data (dissimilarity measures) never exactly fit a tree.

## Neighbor Joining

- ▶ is an algorithmic approach that combines the 4-point condition with averaging to fit data to a tree.
- ▶ is generally considered the best practical distance method for tree construction.
- ▶ is fast —  $\mathcal{O}(n^3)$

## Distance methods

- ▶ are based on 2-sequence comparisons
- ▶ Hamming distance can be improved with a probabilistic model of evolution  
distance = expected no. of mutations per site
- ▶ can be algorithmic (NJ) — faster  
or can use optimality criterion ( $L^2$  fit, ...) — slower

### Weakness:

- ▶ loss of information in summarizing sequence data with pairwise distances (2-marginals)

## Model-based methods.

With a probabilistic model of the mutation process specified, use

Statistical Frameworks:

- ▶ **Maximum Likelihood** – find the parameters  $\Theta$  (especially the tree) that maximize the likelihood function,  $L(\Theta) = P(\text{data} \mid \Theta)$
- ▶ **Bayesian Methods** – find the posterior distribution on the parameters  $\Theta$  (especially the tree)

**Software:** PAUP\*, Phylip, PAML, phyML, RAxML, Garli, Mr. Bayes, etc.

A glimpse into the difficulties of ML:

### Computational challenges

How to search tree space?

For a fixed tree, how to optimize  $L$ ?

multiple maxima? local maxima?

### Statistical challenges

Model specification (too many parameters, too few?)

Model misspecification

Model-based methods require specifying a probabilistic model of the mutation process.

Advantages:

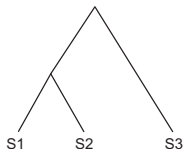
- ▶ Assumptions are explicit (MP – model ?)
- ▶ Inference may use full data (not summary statistic)

# Modeling molecular evolution along a tree $T$ :

Fix an  $n$ -taxon (binary) rooted, leaf-labelled tree  $T$ ,

root = most recent common ancestor

leaves = currently extant taxa



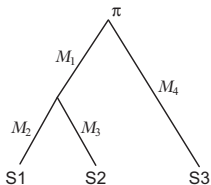
$\kappa$  states at each node,

$\kappa = 4$  (A,C,G,T),

$\kappa = 20$  (proteins)

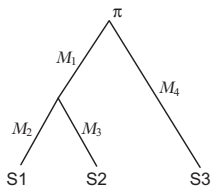
$\kappa = 2$  (R={A,G}, Y={C,T}),

$\kappa = 61$  (codons=triplets of A,C,G,T)



Model parameters =  $\left\{ \begin{array}{l} \text{tree topology} \\ \text{root distribution vector } \pi \\ \text{Markov matrix on each edge } M_e \end{array} \right.$

Model describes evolution at a single site in sequence



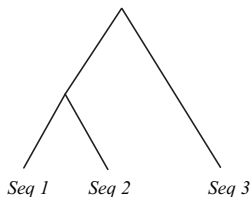
More specifically,

- ▶ States  $1, 2, \dots, \kappa$  ( $A, C, G, T \rightsquigarrow 1, 2, 3, 4$ )
- ▶ State at root given by probabilities  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_\kappa)$ ;  $\sum \pi_i = 1$ .
- ▶ On edge  $e$  Markov matrix  $M_e$  give probs. of state change,

$$M_e(i, j) = P(j \text{ at end} \mid i \text{ at start})$$

This is the **general Markov model** (GM) on the tree  $T$ .

(Other models  $\mathcal{M}$  will appear later....)

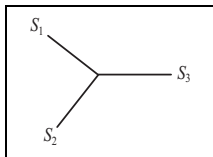
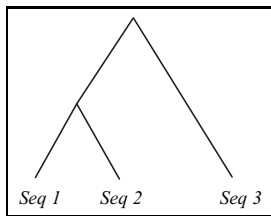
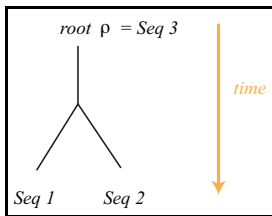


## Note:

- ▶ Multiple sites, assume i.i.d.
- ▶ Data comes only from living taxa at leaves; states at internal nodes are *hidden*. (latent variables)
- ▶ Given state at any node, processes on descending edges are independent. (conditional independence)

- All models discussed today are 'reversible'

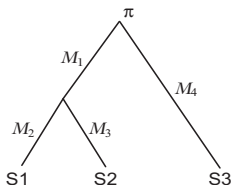
⇒ models on unrooted trees, and thus all three tree models



give rise to the same collection of joint distributions (varieties)

(*hints* at connections to secant varieties ..., re: Sturmfels' talk)

Model parameters  $T, \pi, \{M_e\}$  lead to values for the  
joint distribution of states at leaves

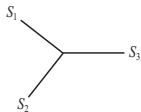


$$p_{ijk} = \sum_{l=1}^{\kappa} \sum_{m=1}^{\kappa} \pi_l M_1(l, m) M_2(m, i) M_3(m, j) M_4(l, k)$$

- $P = (p_{ijk})$  is a  $\kappa \times \kappa \times \kappa$  tensor (table) with entries that
- are polynomial in the stochastic parameters
  - are  $p_{ijk} = P(S1 = i, S2 = j, S3 = k)$
  - can be estimated from data by  $\hat{p}_{ijk}$ .

Using the equivalent unrooted model on 3 taxa, in the case of DNA, we write

$$p_{ijk} = \sum_{f=1}^4 \pi_f M_1(f, i) M_2(f, j) M_3(f, k)$$



Then,  $P = (p_{ijk})$  is a  $4 \times 4 \times 4$  tensor of rank 4, with one summand for each possible state at the internal node.

$$P = (p_{ijk})_A + (p_{ijk})_C + (p_{ijk})_G + (p_{ijk})_T$$

This is precisely the 'mixture of independent distributions' model of Sturmfel's talk. (Here we have four classes.)

In summary,

for a fixed tree  $T$ , we have a polynomial parameterization

$$\Theta \rightarrow \{P\} \subset \Delta$$

given by

$$\theta = \{T, \pi, \{M_e\}\} \mapsto P_\theta = P = (p_{i_1 \dots i_n}).$$

If this map is extended to the complex numbers, then the closure of the image is the *phylogenetic variety*  $V_T$ ,

$$\Theta_{\text{complex}} \rightarrow \text{Im}(\Theta_{\text{complex}}) \subset V_T$$

The parameterized phylogenetic variety  $V_T$  also has an *implicit description*, as the zero set of polynomials in the *phylogenetic ideal*  $I_T$ .

Thus,

$$V_T = Z(I_T),$$

and polynomials in  $I_T$  are known as *phylogenetic invariants*.

$$f \in I_T \iff f(P) = 0 \text{ for all } P = \phi_T(\pi, \{M_e\})$$

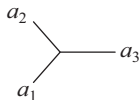
Studying phylogenetic ideals and varieties has led to some interesting results:

- ▶ formulating the ML problem as a *constrained optimization* problem
- ▶ understanding the *geometry* for group-based models
- ▶ connections to *higher secant varieties*
- ▶ *identifiability* results
- ▶ *inference of missing data/formulas for parameter estimation* in more complicated models

# Salmon background

3 taxa, GM,  $\kappa = 4$  (DNA)

(Equivalently, the 'conditional independence' model)



The phylogenetic variety is  $V_T = \text{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$ .

The phylogenetic ideal  $I_T$

- contains the trivial invariant:  $\sum p_{ijk} - 1$
- has no homogeneous invariants of degree  $< 5$ .
- has a 1728-dim space of *all* quintics that can be explicitly constructed
- is not generated by quintics
- = ????

Main result for  $\kappa > 2$ :

**Theorem** (A., Rhodes 2004): For any  $\kappa$ , given all invariants associated to the 3-taxon tree, we can explicitly construct set-theoretic defining polynomials for  $V_T$  for GM model on any binary tree  $T$ .

**Theorem** (Draisma-Kuttler 2008): scheme-theoretic version

In the case of DNA, ( $\kappa = 4$ ), since we still do not know generators of

$$I_{T_3} = I(\text{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)) = ??,$$

there is an opportunity to earn ...

*Reward:* Smoked Copper River Salmon (personally caught for you...)



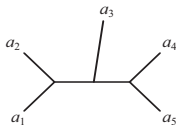
at work for your prize ...

This for a correct solution to

*Problem:* Determine the ideal defining  $\text{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$ .

## Invariants for GM model with 2 states

↔ ranks of flattenings of  $P$  according to topological features in the tree (edges)



Example: For GM,  $\kappa = 2$ ,

The joint distribution tensor  $P$  is  $2 \times 2 \times 2 \times 2 \times 2$ .

$P$  has two natural flattenings according to splits in the tree:

$$a_1 a_2 \mid a_3 a_4 a_5, \text{ and } a_1 a_2 a_3 \mid a_4 a_5.$$

The corresponding flattenings are

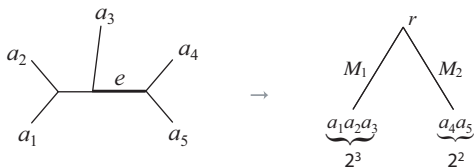
$$\begin{pmatrix} P_{00000} & P_{00001} & P_{00010} & P_{00011} & P_{00100} & P_{00101} & P_{00110} & P_{00111} \\ P_{01000} & P_{01001} & P_{01010} & P_{01011} & P_{01100} & P_{01101} & P_{01110} & P_{01111} \\ P_{10000} & P_{10001} & P_{10010} & P_{10011} & P_{10100} & P_{10101} & P_{10110} & P_{10111} \\ P_{11000} & P_{11001} & P_{11010} & P_{11011} & P_{11100} & P_{11101} & P_{11110} & P_{11111} \end{pmatrix}$$

and

$$\begin{pmatrix} P_{00000} & P_{00001} & P_{00010} & P_{00011} \\ P_{00100} & P_{00101} & P_{00110} & P_{00111} \\ P_{01000} & P_{01001} & P_{01010} & P_{01011} \\ P_{01100} & P_{01101} & P_{01110} & P_{01111} \\ P_{10000} & P_{10001} & P_{10010} & P_{10011} \\ P_{10100} & P_{10101} & P_{10110} & P_{10111} \\ P_{11000} & P_{11001} & P_{11010} & P_{11011} \\ P_{11100} & P_{11101} & P_{11110} & P_{11111} \end{pmatrix}.$$

For this 5-leaf tree,  $I_T$  contains all  $3 \times 3$  minors of these two matrices.  
(That is, these matrices have rank  $\leq 2$ .)

*N.B.* focusing on edge  $e$  leads to a '*simpler*' graphical model:



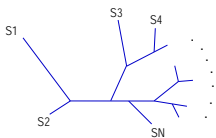
for  $M_1$ , a  $2 \times 2^3$  matrix  
 $M_2$ , a  $2 \times 2^2$  matrix

5-dim  $2 \times \dots \times 2$  tensor  $P \rightarrow 2^3 \times 2^2$  matrix  $\text{Flat}_e(P)$

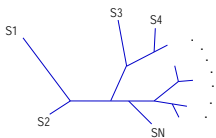
$$\text{Flat}_e(P) = M_1^T \text{diag}(\pi_r) M_2$$

Since  $\text{diag}(\pi)$  is  $2 \times 2$ , this flattening has rank at most 2.

**Theorem** (2004): For  $\kappa = 2$ , any binary  $T$ , the phylogenetic ideal  $I_T$  for the GM model is generated by all  $3 \times 3$  minors of all matrix flattenings of  $P$  on edges of  $T$ .



**Theorem** (2004): For  $\kappa = 2$ , any binary  $T$ , the phylogenetic ideal  $I_T$  for the GM model is generated by all  $3 \times 3$  minors of all matrix flattenings of  $P$  on edges of  $T$ .



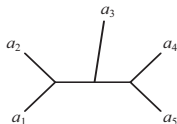
This theorem is useful in surprising ways ...

## Parsimony-informative models:

- ▶ Variants of standard Markov substitution models on trees where *only* parsimony-informative patterns are observed
- ▶ Useful for phenotypic datasets — acquisition bias prevents appropriate sampling of non-informative character patterns (e.g., all equal, all different, ...)
- ▶ Model proposed by P. Lewis (2001) omits constant patterns.  
Model of Ronquist–Huelsenbeck (2004?) omits parsimony-noninformative patterns;  
used for combined analysis of sequence and morphological dataset by Nylander–Ronquist–Huelsenbeck–Nieves-Aldrey (2004)

2-state General Markov model, with only parsimony-informative characters observed

Parameters: Tree,  $2 \times 2$  Markov matrix on each edge, arbitrary root distribution



Example:

The edge flattening of  $P$  corresponding to the split  $a_1 a_2 \mid a_3 a_4 a_5$  is

$$\begin{pmatrix} \mathbf{q}00000 & \mathbf{q}00001 & \mathbf{q}00010 & q00011 & \mathbf{q}00100 & q00101 & q00110 & q00111 \\ \mathbf{q}01000 & q01001 & q01010 & q01011 & q01100 & q01101 & q01110 & \mathbf{q}01111 \\ \mathbf{q}10000 & q10001 & q10010 & q10011 & q10100 & q10101 & q10110 & \mathbf{q}10111 \\ q11000 & q11001 & q11010 & \mathbf{q}11011 & q11100 & \mathbf{q}11101 & \mathbf{q}11110 & \mathbf{q}11111 \end{pmatrix}$$

However, we know only *some* of these entries, rescaled by an unknown factor. *Blue* entries are unknown.

Judicious choices of  $3 \times 3$  minors plus knowledge of GM invariants  
( $3 \times 3$  minors of flattenings vanish)

*allows for determination of unknown entries,*

provided certain  $2 \times 2$  minors don't vanish.

## Determination of unknown entries:

E.g., consider the  $3 \times 3$  minor

$$\begin{vmatrix} q_{01001} & q_{01010} & q_{01011} \\ q_{10001} & q_{10010} & q_{10011} \\ q_{11001} & q_{11010} & \mathbf{q_{11011}} \end{vmatrix} = 0,$$

Expanding the determinant in cofactors by the last column we have

$$q_{01011} \begin{vmatrix} q_{10001} & q_{10010} \\ q_{11001} & q_{11010} \end{vmatrix} - q_{10011} \begin{vmatrix} q_{01001} & q_{01010} \\ q_{11001} & q_{11010} \end{vmatrix} + \mathbf{q_{11011}} \begin{vmatrix} q_{01001} & q_{01010} \\ q_{10001} & q_{10010} \end{vmatrix} = 0$$

Thus provided

$$\begin{vmatrix} q_{01001} & q_{01010} \\ q_{10001} & q_{10010} \end{vmatrix} \neq 0$$

we can determine  $\mathbf{q_{11011}}$  from other  $q_i$  where  $\mathbf{i}$  can be observed.

For 5-taxon trees, enough  $2 \times 2$  minors may be zero to defeat this approach.

However, for larger trees, this approach can be pushed through (also for  $\text{GMk}_{\text{pars-inf}}$  models) yielding....

**Theorem** (2008, A., Holder, Rhodes) Under biologically reasonable assumptions (all entries of  $\pi$  non-zero,  $M_e$  non-singular), parameters  $T, \pi, \{M_e\}$  for the  $\text{GMk}_{\text{pars-inf}}$  model are *identifiable* from the joint distribution of parsimony-informative probabilities.

For the tree parameter, we require  $n \geq 8$  leaves.

For the numerical parameters, if the tree is known, we require  
 $n \geq 7$  leaves.

(Identifiability is necessary for statistical consistency of inference)

# Models of sequence evolution

- ▶ **general Markov model (GM):**

- ▶ arbitrary root distribution
- ▶ arbitrary Markov matrices  $M_e$  on each edge

↪ *Problem:* too many parameters for inference

- ▶ **continuous time models (GTR):**

- ▶ very popular with biologists
- ▶ substitutions occur through a continuous time process, rates are parameters (rate matrix  $Q$ )
- ▶ one parameter per edge – time  $t_e$  of descent
- ▶ Markov matrix on each edge is  $M_e = \exp(Qt_e)$
- ▶ by reducing the number of parameters, avoid overfitting the data

↪ *Comment:* strong assumption of commonality of mutation process along all edges of the tree (too few parameters ?)

## Models incorporating more biological assumptions or complexity

### ▶ Mixture models

- a step towards removing the 'identically distributed' assumption for multiple sites
- Each site belongs to one of  $c$  classes (eg. — if  $c = 3$ , fast, medium, slow), and each class has own set of numerical parameters, but all share the same tree parameter

### Eg. Invariable sites models

these models include a class of *Invariable* sites, (GTR+I, GM+I)

Invariable sites are constrained, and not free to vary.

The proportion of Invariable sites  $p_{inv}$  is unknown, and must be estimated.

## Models incorporating more biological assumptions or complexity

### ▶ Mixture models

- a step towards removing the 'identically distributed' assumption for multiple sites
- Each site belongs to one of  $c$  classes (eg. — if  $c = 3$ , fast, medium, slow), and each class has own set of numerical parameters, but all share the same tree parameter

### Eg. Invariable sites models

these models include a class of *Invariable* sites, (GTR+I, GM+I)

Invariable sites are constrained, and not free to vary.

The proportion of Invariable sites  $p_{inv}$  is unknown, and must be estimated.

**Elaborations of interest to biologists:** GTR+ $\Gamma$ , GTR+I+ $\Gamma$ , tree mixtures, covarion ('class-switching') models

## Mixtures, covarion models, non-identifiability ...

- ▶ Modeling evolution as a heterogeneous process is desirable
  - ▶ *heterotachy* (across-sites, across-lineages) — helps model biological phenomena like
    - redundancy in the genetic code;
    - speeding up or slowing down of evolution in distant regions of large phylogenies;
    - different evolutionary pressures on multi-gene data sets
  - ▶ other models attempt to describe lateral gene transfer, incomplete lineage sorting, changing base distributions, ...
- ▶ tremendous amount of recent interest

## Identifiability of model parameters

If  $\mathcal{T}$  denotes tree space and  $\mathcal{M}$  any choice of model, then  $\mathcal{D}_{\mathcal{M}}$  is the image of the parameterization map(s)

$$\begin{aligned}\phi_{\mathcal{M}} : \bigcup_{T \in \mathcal{T}} (T, S_T) &\longrightarrow [0, 1]^{\kappa^n} \\ (T, s_T) &\longmapsto P = \phi_{\mathcal{M}, T}(s_T)\end{aligned}$$

**Question.** Suppose  $P$  is a joint distribution arising from model parameters  $(T, s_T)$  for  $\mathcal{M}$ . Can we *identify*  $(T, s_T)$  from its image?

i.e. Is the map  $\phi_{\mathcal{M}}$  above invertible?

**Identifiability is needed for statistical consistency of inference.**

## First considerations:

How many taxa at a time should we look at?

Strepsiptera	AAGCTCATTAAATCGCTTTGGTTCCTTAGATAGTTGGAT	...	$\infty$
Aedes	AGGCTCAGTATAACACTATAATTTACAAGATCATTGGAT	...	$\infty$
Drosophila	AGGCTCATTATATCATTATGGTTCCTTAGATCGTTGGAT	...	$\infty$
Flea	TGGCTCATTATATCATTATGGTTCATTAGATCGTTGGAT	...	$\infty$
Meloe	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT	...	$\infty$
Tenebrio	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT	...	$\infty$

More precisely, how much information do  $k$ -marginals hold about  $P$ ?

## First considerations:

How many taxa at a time should we look at?

→	Strepsiptera	AAGCTCATTAAATCGCTTTGGTTCCTTAGATAGTTGGAT	...	∞
	Aedes	AGGCTCAGTATAACACTATAATTTACAAGATCATTGGAT	...	∞
	Drosophila	AGGCTCATTATATCATTATGGTTCCTTAGATCGTTGGAT	...	∞
	Flea	TGGCTCATTATATCATTATGGTTCATTAGATCGTTGGAT	...	∞
	Meloe	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT	...	∞
	Tenebrio	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT	...	∞

More precisely, how much information do  $k$ -marginals hold about  $P$ ?

For a stationary model (eg. GTR), from **1-taxon** marginalizations we can recover the base distribution  $\pi$ .

## First considerations:

How many taxa at a time should we look at?

	Strepsiptera	AAGCTCATTAAATCGCTTTGGTTCCTTAGATAGTTGGAT	...	$\infty$
	Aedes	AGGCTCAGTATAACACTATAATTTACAAGATCATTGGAT	...	$\infty$
→	Drosophila	AGGCTCATTATATCATTATGGTTCCTTAGATCGTTGGAT	...	$\infty$
	Flea	TGGCTCATTATATCATTATGGTTCATTAGATCGTTGGAT	...	$\infty$
→	Meloe	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT	...	$\infty$
	Tenebrio	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT	...	$\infty$

More precisely, how much information do  $k$ -marginals hold about  $P$ ?

For a stationary model (eg. GTR), from **1-taxon** marginalizations we can recover the base distribution  $\pi$ .

For a model with a distance (eg. GTR, GM, group-based), from **2-taxon** marginalizations we can recover the tree  $T$ .

For those models with distances,

How many taxa at a time should we look at?

→	<b>Strepsiptera</b>	AAGCTCATTAAATCGCTTTGGTTCCTTAGATAGTTGGAT	...	∞
	<b>Aedes</b>	AGGCTCAGTATAACACTATAATTTACAAGATCATTGGAT	...	∞
→	<b>Drosophila</b>	AGGCTCATTATATCATTATGGTTCCTTAGATCGTTGGAT	...	∞
→	<b>Flea</b>	TGGCTCATTATATCATTATGGTTCATTAGATCGTTGGAT	...	∞
	<b>Meloe</b>	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT	...	∞
	<b>Tenebrio</b>	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT	...	∞

From **3-taxon** marginalizations we can recover the edge transition matrices  $\{M_e\}$ . (Chang 1996)

However, some models don't have distances.

→ the standard approach to tree identifiability fails

Baake 1998:

Constructs  $P_1$  and  $P_2$  on different 4-taxon trees with same 2-marginals for a mixture model with invariable sites (GM+I).

Implications:

2-marginals not enough to identify  $T$  for mixtures

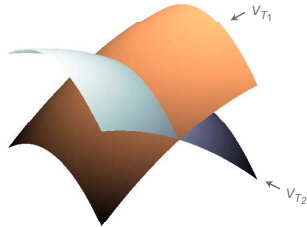
such models do not have distances

One solution:

invariants, geometry, and generic identifiability of tree parameter

Fixing  $\mathcal{M}$ , then  $V_T$  are geometric objects

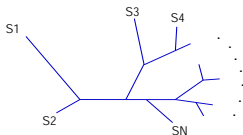
$\rightsquigarrow$  irreducible, with positive dimensions



Moreover,

if the intersection  $V_{T_1} \cap V_{T_2}$  is of lower dimension, then the  
*tree is identifiable for generic parameters.*

**Theorem** (2006): Trees are identifiable for generic parameters for a generalized GM model with



$M_{internal}$  of size  $\lambda \times \lambda$

$M_{pendant}$  of size  $\lambda \times \kappa$ ,  $\lambda < \kappa^2$

Proof: Construction of only a few invariants in  $I_T$ , but enough to show  $V_T \neq V_{T'}$  if  $T \neq T'$ .

Invariants come from rank conditions on flattening along edges.

Flat $_e(P)$  has rank  $\begin{cases} \leq \lambda & \text{if } e \text{ is an edge in the true tree;} \\ > \lambda & \text{if } e \text{ not in true tree.} \end{cases}$

Using this algebraic perspective,

we obtain results on mixture models of biological interest on a single tree with 'small' number of classes (after specialization).

'Small' means 'less than the number of states.'

For DNA models, this proves tree is generically identifiable for GTR+I, GTR+d $\Gamma$ (3), GTR+GTR+GTR, GM+GM+GM, covarion model with 3 rate classes, ...

For protein models, .... up to 19-class mixture models, etc.

For codon models, .... up to 60-class mixture models, etc.

Other algebraic methods have led to proofs of the identifiability of generic parameters for

- ▶ covarion models

Numerical parameters for very general covarion models from 7-marginals. (2008) A., Rhodes

- ▶ parameters in random graph model, HMMs, Mixtures of products of nonparametric distributions. (2008) A., Matias, Rhodes

.....

# Open problem

For the most commonly-used model in phylogenetic inference,

GTR+I+ $\Gamma$ ,

*it has not been proved* that parameters are identifiable.

## History:

— (1990s): GTR is identifiable

Rogers (2001): flawed proof for GTR+I+ $\Gamma$

A-Rhodes (2004-7): GTR+I

A-Ané-Rhodes (2008): GTR+ $\Gamma$  (from 3-marginals)

Steel (2008): For GTR+ $\Gamma$ , neither the tree nor  $\alpha$ -shape parameter are identifiable from 2-marginals

Thank you!