The Identifiability of Covarion Models

John A. Rhodes

University of Alaska

Fairbanks

December 18, 2007

Isaac Newton Institute

Joint work with Elizabeth Allman,

Thanks to the INI,

and numerous PLG participants.

I: Covarion models:

Simple models of sequence evolution (JC, Kimura, HKY, GTR)

assume evolutionary process is

- homogeneous along sequence
- homogeneous through time

Both assumptions are likely violated.

Models with additional realism:

Inhomogeneity along sequence $\rightsquigarrow$ across-site rate-variation models

$\mathcal{M}$+I, $\mathcal{M}$+Γ, $\mathcal{M}$+Γ+I

Inhomogeneity through time $\rightsquigarrow$ Covarion models

Nucleotide example:

2 rate-classes, 'bases' for each class

$$\mathsf{A}^1, \mathsf{C}^1, \mathsf{G}^1, \mathsf{T}^1, \qquad \mathsf{A}^2, \mathsf{C}^2, \mathsf{G}^2, \mathsf{T}^2$$

GTR rate-matrix (4×4): Q

Switching parameters: $s_1, s_2 > 0$

Rate-class scaling factor: $1 > r \geq 0$

Create an 8×8 rate matrix

$$R = \begin{pmatrix} Q - s_1 I & s_1 I \\ s_2 I & rQ - s_2 I \end{pmatrix}$$

$$
\begin{array}{rcc}
 & A^1, C^1, G^1, T^1 & A^2, C^2, G^2, T^2 \\
 & \downarrow & \downarrow \\
A^1, C^1, G^1, T^1 \rightarrow & Q_1 - s_1 I & s_1 I \\
A^2, C^2, G^2, T^2 \rightarrow & s_2 I & Q_2 - s_2 I
\end{array}
$$

Within-class substitution rates: $Q_2 = rQ_1$

Between-class switching rates: s_1, s_2

The combined process is time-reversible, stationary.

Data is *only* A, C, T, G , no superscripts

All rate-class information is hidden.

$$A^1, A^2 \rightsquigarrow A$$

$$C^1, C^2 \rightsquigarrow C$$

$$G^1, G^2 \rightsquigarrow G$$

$$T^1, T^2 \rightsquigarrow T$$

Generalizations:

— κ observable states, c classes, $c\kappa \times c\kappa$ rate matrices

(e.g., protein models $\kappa = 20$, codon models $\kappa = 61$)

— Within-class matrices need *not* be scaled versions of a single Q,

as long as entire process is time-reversible

— Switching only between certain classes

(e.g., incorporate across-site variation)

Tuffley - Steel 1998: mathematical formalization (2 classes, 1 of which invariable)

– only partially captures covarion idea of Fitch - Markowitz (1970)

Galtier 2001: 4-class software implementation

Huelsenbeck 2002: 2-class covarion, but with g rates-across-sites classes $\rightsquigarrow 2g$ classes

Galtier - Jean-Marie 2004: theoretical observations on eigenvectors of scaled covarion models

Guindon - Rodrigo - Dyer - Hulesenbeck 2004: codon model with classes representing selection regimes

Reviews, with further generalization:

- Gascuel - Guindon, 2007 in "Reconstructing Evolution,"
- Wang - Spencer - Susko - Roger, 2007, MBE

Talks Thursday:

- Roger
- Whelan

Other names:

covariotide

covarion-like

site-specific rate variation (SSRV)

Markov-modulated Markov Process (MMM)

II: Identifiability:

A model is **identifiable** if

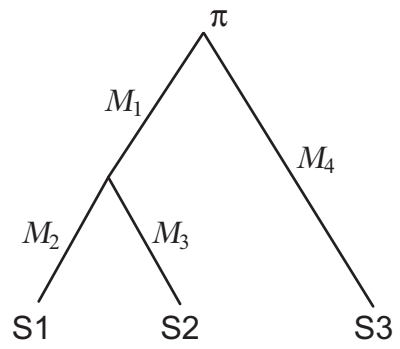
given an exact distribution of site-patterns arising from the model,

"infinite amount of perfect data,"

we can determine all model parameters.

Identifiability is necessary for statistical consistency of inference.

For covarion models,



$$\text{parameters} = \begin{cases} \text{Tree topology } T \\ \text{Substitution rates } Q_i \text{ for class } i = 1, 2, \ldots, c \\ \text{Switching rates } S \text{ between classes} \\ \text{Stationary distribution } \boldsymbol{\mu} \text{ for full process} \\ \text{Edge lengths on tree} \end{cases}$$

It is not simple to see that *any* of these are identifiable.

III: Results:

Theorem: (Allman, R—, 2006)

For generic parameters of the covarion model (and mixture models), the tree topology is identifiable provided $c < \kappa$,

i.e., if fewer classes than observable states.

Proof: Uses phylogenetic invariants

New:

Theorem: (Allman, R—)

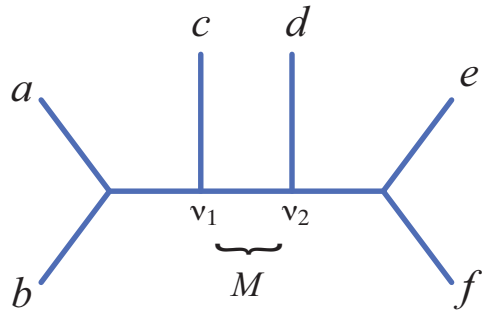
For generic parameters of the covarion model, if the tree topology is known and has at least 7 leaves, then all other parameters are identifiable provided

- $c \leq \kappa$, $\quad$ ($c < \kappa$ for S. Whelan)
- The switching process is irreducible (possible to switch from any class to any other class, through multiple steps)
- the entire process is time-reversible

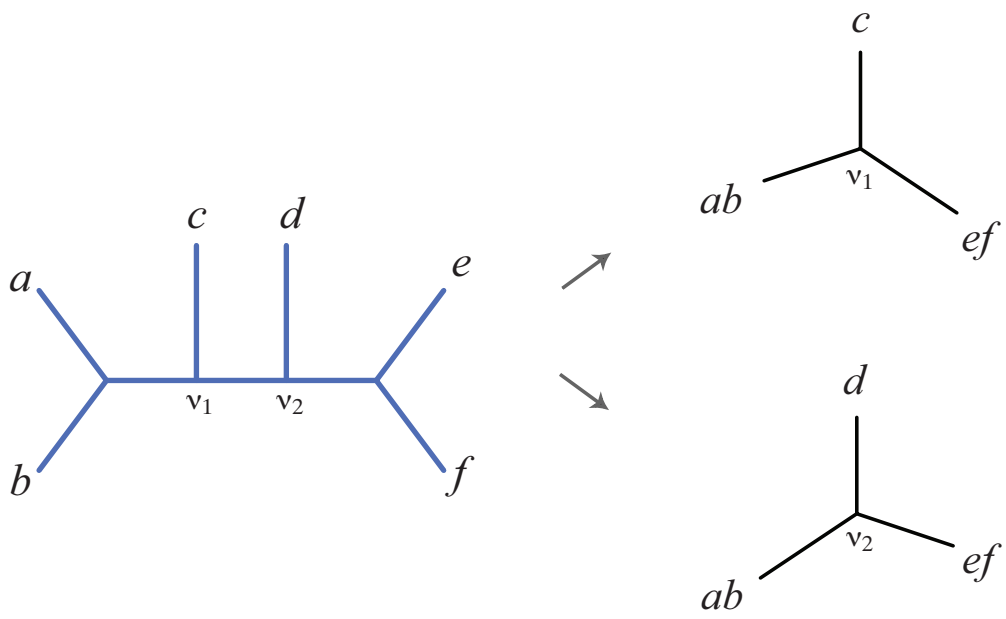
Note: $Q_i = r_i Q$ is *not* necessary.

Sketch of Proof:

With 7 leaves, T has a sub-tree



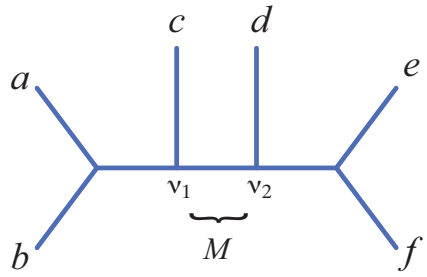
Goal is to identify $M = \exp(Rt)$ deep in tree



A theorem of J. Kruskal (1977) gives algebraic conditions for the identifiability (up to permutations) of general 3-leaf models.

Using Kruskal, we can identify M for the general model

– but must painfully show technical algebraic conditions are met –



$M = \exp(Rt)$ is not an algebraic function of its variables.

But the rows and columns of $M = \exp(Rt)$ are only determined up to permutation, further argument is needed to order correctly

— use special form of R,

pattern of zero/non-zero entries

irreducibility of switching process

time reversibility

Then matrix logarithm recovers R.

Finally, use R and 2-taxa marginalizations to get distances between leaves, which determines edge lengths

Final comments:

Identifiability of covarion model is known only for generic parameters

There may be parameters for which model is *not* identifiable.

but such cases must be 'rare' (of measure zero)

Generic identifiability is common for complex statistical models with hidden variables.

Kruskal's Theorem:

Consider a 3-leaf star model with R hidden states at the central root r and any number of states at the leaves a, b, c.

Let $\boldsymbol{\nu} \in \mathbb{R}^R$ be a root distribution, and $M_{ra},\ M_{rb},\ M_{rc}$ Markov matrices on the edges.

Define $I_a = \max\{k \mid$ *every* set of k rows of M_{ra} is independent$\}$, and I_b, I_c, similarly.

Theorem: If all entries of $\boldsymbol{\nu}$ are non-zero and

$$I_a + I_b + I_c \geq 2R + 2,$$

then the model's probability distribution uniquely determines $\boldsymbol{\nu}$, M_{ra}, M_{rb}, M_{rc}, up to some permutation.

I.e, for some unknown permutation P we may find

$$\boldsymbol{\nu}P^T,\ PM_{ra},\ PM_{rb},\ PM_{rc}.$$