Systematics of phylogenetic models

. . . symmetry groups and Markov invariants

Peter Jarvis

School of Mathematics and Physics

University of Tasmania

- Models within models?

  $\cdots$ Multiplicative closure and the trouble with GTR

- Examples

  $\cdots$ symmetrically embedded $2 \hookrightarrow K$ state models

- Markov invariants for phylogenetic models

- $2 \hookrightarrow 3$ case and simulations

- Conclusions

Phylogenetics

- Taxa are represented by DNA sequences

  ACGTTGAACTGG$\cdots$

  (or RYRYYRRRYYRR$\cdots$)

- Genetic information content is sparse, so we can talk about *probabilities* (relative frequencies)

  p_A, p_C, p_G, p_T with $p_A + p_C + p_G + p_T = 1$

  (or p_0, p_1 with $p_0 + p_1 = 1$)

- Mutations are mostly 'neutral', so these probabilities are subject to *random changes* under a Markov process – just given by matrix multiplication:

$$\begin{pmatrix} p_0 \\ p_1 \end{pmatrix} \to \begin{pmatrix} m_1 & m_2 \\ m_3 & m_4 \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \end{pmatrix} \quad \Leftrightarrow \quad p \to M\,p$$

Branching?

- We form an array of patterns *(relative frequencies)*, from the base sequences of an alignment of genes from different species, for example with two or three species

  ACGTTGAACTGG$\cdots$ ACGTTGAACTGG$\cdots$
  AAGTCGAACACG$\cdots$ AAGTCGAACACG$\cdots$
  AATTCGATCAGG$\cdots$

- Thus we have 16 or 64 or $4^{\#\text{ of leaves}}$ patterns respectively:

$$(p_{AA}, p_{AC}, p_{AG}, p_{AT}; p_{CA}, \cdots; \cdots, p_{TG}, p_{TT})$$
$$(p_{AAA}, p_{AAC}, p_{AAG}, p_{AAT}; p_{ACA}, \cdots; \cdots, p_{TTG}, p_{TTT})$$

or with two letters

$$(p_{00}, p_{01}, p_{10}, p_{11})$$

or

$$(p_{000}, p_{001}, \cdots; \cdots p_{110}, p_{111}).$$

Models within models?

- Standard model for phylogenetic process on a tree assumes Markov transition matrices on each edge.
- For heterogeneous models – e.g. differing transition rates & base compositions, apart from edge lengths – as well as mosaic models or even mixture models, *multiplicative closure* of Markov matrices is mandatory for consistency.
- In continuous time models with a rate matrix, $M = \exp tQ$ and exponentials combine via the BCH formula,

$$M_1 M_2 = e^{Q_1} e^{Q_2} = e^{Q_1 + Q_2 + \frac{1}{2}[Q_1, Q_2] + \frac{1}{12}[[Q_1, Q_2], Q_2] \pm \cdots}$$

  where $[A, B] = AB - BA$, so closure is guaranteed if the Q's generate a *Lie algebra*.
- Closure is also implicated under incomplete sampling.

- Kimura models (group $\cong GL(1) \times GL(1) \times GL(1)$):

$$Q = \begin{pmatrix} -(\alpha+\beta+\gamma) & \gamma & \alpha & \beta \\ \gamma & -(\alpha+\beta+\gamma) & \beta & \alpha \\ \alpha & \beta & -(\alpha+\beta+\gamma) & \gamma \\ \beta & \alpha & \gamma & -(\alpha+\beta+\gamma) \end{pmatrix}$$

$$\equiv \alpha \begin{pmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} + \beta \begin{pmatrix} -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix} + \gamma \begin{pmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

$$\equiv \alpha K_\alpha + \beta K_\beta + \gamma K_\gamma$$

This is a 3D maximal *abelian*, (Cartan) Lie algebra – the K_X can be simultaneously diagonalised via the Hadamard transformation (Bashford, PDJ, Sumner, Steel 2004).

- Felsenstein model: (group $\cong (\times^3 GL(1)) \rtimes GL(1)$):

$$Q = \begin{pmatrix} -\alpha(\pi_C+\pi_G+\pi_T) & \alpha\pi_C & \alpha\pi_G & \alpha\pi_T \\ \alpha\pi_A & -\alpha(\pi_A+\pi_G+\pi_T) & \alpha\pi_G & \alpha\pi_T \\ \alpha\pi_A & \alpha\pi_C & -\alpha(\pi_A+\pi_C+\pi_T) & \alpha\pi_T \\ \alpha\pi_A & \alpha\pi_C & \alpha\pi_G & -\alpha(\pi_A+\pi_C+\pi_G) \end{pmatrix}$$

$$\equiv \alpha \left[\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{pmatrix} - \begin{pmatrix} 1-\pi_A & 0 & 0 & 0 \\ 0 & 1-\pi_C & 0 & 0 \\ 0 & 0 & 1-\pi_G & 0 \\ 0 & 0 & 0 & 1-\pi_T \end{pmatrix} \right]$$

$$= \alpha \left[\pi_A \begin{pmatrix} 1 & 0 & 0 & -1 \\ 1 & 0 & 0 & -1 \\ 1 & 0 & 0 & -1 \\ 1 & 0 & 0 & -1 \end{pmatrix} + \pi_C \begin{pmatrix} 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & -1 \end{pmatrix} + \pi_G \begin{pmatrix} 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} + \begin{pmatrix} -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \right]$$

$$\equiv \alpha\pi_A F_A + \alpha\pi_C F_C + \alpha\pi_G F_G + \alpha F_\circ.$$

Thus we have a *nonabelian*, 4D Lie algebra with $[F_\circ, F_X] = F_X$ and $[F_X, F_Y] = 0$ for all X, Y.

A problem with GTR matrices?

- GTR rate matrices are required to satisfy

$$\Pi Q = Q^\top \Pi$$

where Π is the diagonal matrix of stationary probabilities diag$(\pi_A, \pi_C, \pi_G, \pi_T)$.

- Unfortunately however

$$\Pi[Q_1, Q_2] = \Pi(Q_1 Q_2 - Q_2 Q_1) = (Q_1^\top Q_2^\top - Q_2^\top Q_1^\top)\Pi$$

$$\therefore \quad \Pi[Q_1, Q_2] = -[Q_1, Q_2]^\top \Pi \qquad \textbf{d'oh}$$

- Huelsenbeck, Larget & Alfaro (2004) identified "203 sub-models of GTR $\cdots$ including 'Aunt Emily's' favourite $\cdots$".

- $\cdots$ note – the *general* Markov model (*12* parameters) *is* multiplicatively closed (group $\cong GL_1(4)$) !!

Symmetrically embedded 2 state models

- Consider general 2 state model

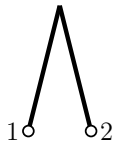
$$Q = \begin{pmatrix} -\alpha & \beta \\ \alpha & -\beta \end{pmatrix} = \alpha \begin{pmatrix} -1 & 0 \\ 1 & 0 \end{pmatrix} + \beta \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \equiv \alpha L_\alpha + \beta L_\beta$$

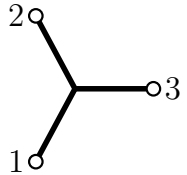
- Consider a derived action, on the distribution (p, q) extended to 3 characters $(p^2, pq = qp, q^2) \equiv (p_1, \frac{1}{2}p_2, p_3)$

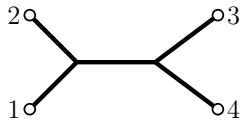
$$Q_3 = \begin{pmatrix} -2\alpha & \beta & 0 \\ 2\alpha & -\alpha-\beta & 2\beta \\ 0 & \alpha & -2\beta \end{pmatrix} = \alpha \begin{pmatrix} -2 & 0 & 0 \\ 2 & -1 & 0 \\ 0 & 1 & 0 \end{pmatrix} + \beta \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & 2 \\ 0 & 0 & -2 \end{pmatrix} \equiv \alpha L_\alpha^{(3)} + \beta L_\beta^{(3)}$$

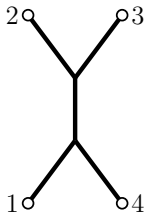
- The Lie algebra $[L_\alpha, L_\beta] = L_\alpha - L_\beta$ is preserved , i.e. L_α^3 and L_β^3 have the same commutation relations and provide an embedding of the 2 state model into the general 3 state model.
- Generalistion to $2 \hookrightarrow 4, \cdots, 2 \hookrightarrow K$ embedded models.

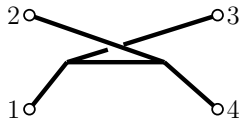
Markov invariants

- Markov invariants (Sumner, Charleston, Jermiin, PDJ 2008) are polynomials of fixed degree in the phylogenetic pattern frequencies (or divergence arrays) for a given number of taxa, which have natural covariance properties with respect to the action of the Markov matrices on leaf edges
- Examples:
- Det (degree K, $K!$ terms, on 

)
- Tangle $\mathcal{T}$, ($K=2$, degree 3, 13 terms, on 

)
- Squangles $\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3$ ($K=4$, degree 5, $\cong 10^5$ terms, on 

, 

, and 

,

  – are phylogenetically informative)

Markov invariants & symmetric embeddings

- Proposition (PDJ & Sumner 2009): for the $2 \hookrightarrow 3$ symmetric embedding and 2 leaves, there are two invariants of quadratic degree, I_{31} and I_{22} such that

$$I_{31}[\,\underset{1}{\circ}\!\!\bigwedge\!\!\underset{2}{\circ}\,] = |M_1||M_2| \cdot (\pi_2\pi_3 + \pi_1\pi_2 + 4\pi_1\pi_3),$$

$$I_{22}[\,\underset{1}{\circ}\!\!\bigwedge\!\!\underset{2}{\circ}\,] = |M_1|^2|M_2|^2 \cdot \tfrac{1}{8}(\pi_2^2 + 8\pi_1\pi_3)$$

$$(\text{c.f.} \qquad \text{Det}[\,\underset{1}{\circ}\!\!\bigwedge\!\!\underset{2}{\circ}\,] = |M_1|^3|M_2|^3 \cdot (\pi_1\pi_2\pi_3) \qquad)$$

where $M = \exp(tQ)$ is the 2×2 model on each edge, with determinants $|M_1| = e^{-t_1}$, $|M_2| = e^{-t_2}$.

Markov invariants & symmetric embeddings II

- Explicitly,

$$I_{31} = 4\big(f_{33} + \tfrac{1}{2}(f_{23} + f_{32}) + \tfrac{1}{4}f_{22}\big) - 4\big(\tfrac{1}{2}f_{12} + \tfrac{1}{2}f_{22} + \tfrac{1}{2}f_{32} + f_{13} + f_{23} + f_{33}\big)\cdot\big(\tfrac{1}{2}f_{21} + \tfrac{1}{2}f_{22} + \tfrac{1}{2}f_{23} + f_{31} + f_{32} + f_{33}\big),$$

$$\begin{aligned} I_{22} = {} & f_{33} + 2\big(f_{33} + \tfrac{1}{2}(f_{23} + f_{32}) + \tfrac{1}{4}f_{22}\big)^2 + \\ & \big(f_{13} + f_{23} + f_{33}\big)\cdot\big(f_{31} + f_{32} + f_{33}\big) \\ & - 2\big(\tfrac{1}{2}f_{12} + \tfrac{1}{2}f_{22} + \tfrac{1}{2}f_{32} + f_{13} + f_{23} + f_{33}\big)\cdot\big(\tfrac{1}{2}f_{23} + f_{33}\big) \\ & - 2\big(\tfrac{1}{2}f_{21} + \tfrac{1}{2}f_{22} + \tfrac{1}{2}f_{23} + f_{31} + f_{32} + f_{33}\big)\cdot\big(\tfrac{1}{2}f_{32} + f_{33}\big). \end{aligned}$$

(c.f. $\mathrm{Det} = f_{11}f_{22}f_{33} + f_{12}f_{23}f_{31} + f_{13}f_{32}f_{21} - f_{11}f_{23}f_{32} - f_{13}f_{22}f_{31} - f_{12}f_{21}f_{33}.$)

Simulations

- Theoretical pattern distribution can be evaluated for various choices of α_1, α_2, t_1, t_2 ($\alpha + \beta = 1$).
- Simulated frequencies give reconstructed I_{31}, I_{22} and Det, compared with theoretical values (taking root frequencies $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$).
- $N = 1000$, $\alpha_1 = .4$, $t_1 = .2$; $\alpha_2 = .5$, $t_2 = .8$:

(Theory)											
$2.542(\times 10^{-1})$	I_{31}	2.430	2.429	2.050	2.576	2.538	2.753	2.423	2.671	2.670	2.035
$1.691(\times 10^{-2})$	I_{22}	1.806	1.659	0.751	2.020	1.982	1.957	1.359	1.876	1.804	1.173
$1.844(\times 10^{-3})$	Det	2.141	1.830	0.336	2.639	2.419	1.851	1.208	1.636	1.983	1.089

- $N = 100$, $\alpha_1 = .5$, $t_1 = .1$; $\alpha_2 = .5$, $t_2 = 1.0$:

(Theory)											
$2.219(\times 10^{-1})$	I_{31}	2.180	2.299	1.552	2.120	2.147	2.000	2.887	2.145	1.660	2.600
$1.385(\times 10^{-2})$	I_{22}	-0.059	0.855	1.715	2.505	2.176	1.581	1.531	1.991	0.781	0.881
$1.366(\times 10^{-3})$	Det	-1.683	0.456	1.544	3.408	2.233	2.190	1.588	2.700	1.420	1.082

- Apparently the lower weight invariants behave better, $\Delta I_{31} < \Delta I_{22} < \Delta Det$

Conclusions

- Multiplicative closure is mandatory for consistent model building and a simple interpretation.
- Many models are multiplicative because the Q's are either abelian (K3ST, 3 parameters) or maximal (GMM, 12 parameters)
- Felsenstein model (4 parameters) is nonabelian
- The $2 \hookrightarrow K$ symmetric embeddings are a class of models with underlying Lie algebra $GL_1(2)$ (2 parameters).

Conclusions- ctd

- Symmetric embeddings also exist for other cases but in specific dimensions, e.g. $3 \hookrightarrow 10$, $4 \hookrightarrow 10$, $4 \hookrightarrow 20$.
- Phylogenetic models using such rate matrices admit analysis by suitable Markov invariants, enabling inference to be drawn on the basis of arbitrary rate matrices from the entire embedding class, e.g. 12 parameters in the case of $4 \hookrightarrow 20$.
- Simulations in the simplest $2 \hookrightarrow 3$ case confirm the theory and show that the quadratic invariants behave better than the cubic Det invariant.

	D	L	
(2,2,3)	1	2	1
		3	1
		4	1
	2	2	5
		3	14
		4	41
	3	2	9
		3	58
		4	401
	4	2	23
		3	321
		4	5 597
(2,3,4)	1	1	1
		3	1
		4	1
	2	2	8
		3	32
		4	128
	3	2	26
		3	292
		4	3 464
	4	2	100
		3	3 688
		4	158 384

	D	L	
(3,2,6)	1	2	1
		3	1
		4	1
	2	2	-
		3	-
		4	-
	3	2	2
		3	4
		4	8
	4	2	4
		3	31
		4	274
(3,3,10)	1	2	1
		3	1
		4	1
	2	2	-
		3	-
		4	-
	3	2	5
		3	13
		4	41
	4	2	19
		3	338
		4	6 532

	D	L	
(4,2,10)	1	2	1
		3	1
		4	1
	2	2	-
		3	-
		4	-
	3	2	-
		3	-
		4	-
	4	2	2
		3	4
		4	8
(4,3,20)	1	2	1
		3	1
		4	1
	2	2	-
		3	-
		4	-
	3	2	-
		3	-
		4	-
	4	2	2
		3	4
		4	8

Table 1: Enumeration of linearly independent candidate Markov invariants, for small numbers of taxa L, and degrees D up to 4.

The stangle polynomials (3 leaves)

- 2 states: degree 3, in 8 variables (5 terms)

$$\mathcal{T}_2^s = -2\psi_{001}\psi_{001}\psi_{001} + \psi_{000}\psi_{011}\psi_{100} + \psi_{000}\psi_{010}\psi_{101} + \psi_{000}\psi_{001}\psi_{110} - \psi_{000}^2\psi_{111}$$

- 3 states: degree 4, in 27 variables (32 terms)

$$\begin{aligned}\mathcal{T}_3^{(s)} &= \Psi_{000011102220} \\ &= \psi_{012}\psi_{020}\psi_{101}\psi_{200} - \psi_{010}\psi_{022}\psi_{101}\psi_{200} - \psi_{011}\psi_{020}\psi_{102}\psi_{200} \\ &+ \psi_{010}\psi_{021}\psi_{102}\psi_{200} - \psi_{002}\psi_{021}\psi_{110}\psi_{200} + \psi_{001}\psi_{022}\psi_{110}\psi_{200} \\ &+ \psi_{002}\psi_{011}\psi_{120}\psi_{200} - \psi_{001}\psi_{012}\psi_{120}\psi_{200} - \psi_{012}\psi_{020}\psi_{100}\psi_{201} \\ &+ \psi_{010}\psi_{022}\psi_{100}\psi_{201} + \psi_{002}\psi_{020}\psi_{110}\psi_{201} - \psi_{000}\psi_{022}\psi_{110}\psi_{201} \\ &- \psi_{002}\psi_{010}\psi_{120}\psi_{201} + \psi_{000}\psi_{012}\psi_{120}\psi_{201} + \psi_{011}\psi_{020}\psi_{100}\psi_{202} \\ &- \psi_{010}\psi_{021}\psi_{100}\psi_{202} - \psi_{001}\psi_{020}\psi_{110}\psi_{202} + \psi_{000}\psi_{021}\psi_{110}\psi_{202} \\ &+ \psi_{001}\psi_{010}\psi_{120}\psi_{202} - \psi_{000}\psi_{011}\psi_{120}\psi_{202} + \psi_{002}\psi_{021}\psi_{100}\psi_{210} \\ &- \psi_{001}\psi_{022}\psi_{100}\psi_{210} - \psi_{002}\psi_{020}\psi_{101}\psi_{210} + \psi_{000}\psi_{022}\psi_{101}\psi_{210} \\ &+ \psi_{001}\psi_{020}\psi_{102}\psi_{210} - \psi_{000}\psi_{021}\psi_{102}\psi_{210} - \psi_{002}\psi_{011}\psi_{100}\psi_{220} \\ &+ \psi_{001}\psi_{012}\psi_{100}\psi_{220} + \psi_{002}\psi_{010}\psi_{101}\psi_{220} - \psi_{000}\psi_{012}\psi_{101}\psi_{220} \\ &- \psi_{001}\psi_{010}\psi_{102}\psi_{220} + \psi_{000}\psi_{011}\psi_{102}\psi_{220}.\end{aligned}$$

- 4 states: degree 6, in 256 variables (1405 terms)