Lie Markov models

Jeremy Sumner

School of Physical Sciences
University of Tasmania, Australia

Stochastic Modelling Meets Phylogenetics, UTAS, November 2015

The theory of (matrix) Lie groups $\mathcal{G}$ and Lie algebras $\mathcal{L}$

- Consider the *orthogonal group* with $MM^T = \mathbf{1}$

  i. $(M_1M_2)(M_1M_2)^T = M_1(M_2M_2^T)M_1^T = \mathbf{1}$

  ii. $\mathbf{1} = M^{-1}(MM^T)(M^{-1})^T = M^{-1}\mathbf{1}(M^{-1})^T = M^{-1}(M^{-1})^T$

- Consider path $M(t)$ and $M(0) = \mathbf{1}$.

  Tangents $X := \left.\frac{dM(t)}{dt}\right|_0$ satisfy $X + X^T = 0$

- Forms a *Lie algebra* $\mathcal{L}$:

  i. $X + \lambda Y \in \mathcal{L}$

  ii. $[X, Y] := XY - YX \in \mathcal{L}$

- Exponential map: $exp : \mathcal{L} \to \mathcal{G}$

  i.e. $exp(X)exp(X)^T = exp(X)exp(-X) = \mathbf{1}$✓

DNA substitutions modelled as cont-time Markov chain

- Model sequence evolution as a CTMC on nucleotides $\{A, G, C, T\}$
- Two extremes: "All rates are the same" OR "All rates (might be!) different".

$$\begin{pmatrix} * & \alpha & \alpha & \alpha \\ \alpha & * & \alpha & \alpha \\ \alpha & \alpha & * & \alpha \\ \alpha & \alpha & \alpha & * \end{pmatrix} \quad \text{OR?} \quad \begin{pmatrix} * & \alpha & \beta & \gamma \\ \delta & * & \epsilon & \phi \\ \psi & \zeta & * & \varphi \\ \xi & \omega & \sigma & * \end{pmatrix}$$

- What model is best depends on bias-variance tradeoff.
- Lots of molecular data means model complexity has somewhat been driven by computing power.

The GTR model (Tavare 1986)

- Stationary dist: $\pi = (\pi_A, \pi_G, \pi_C, \pi_T)^T$
- Time reversible: rate $A \rightarrow T$ *equals* rate $T \rightarrow A$

$$Q = \begin{pmatrix} * & \pi_A s_1 & \pi_A s_2 & \pi_A s_3 \\ \pi_G s_1 & * & \pi_G s_4 & \pi_G s_5 \\ \pi_C s_2 & \pi_C s_4 & * & \pi_C s_6 \\ \pi_T s_3 & \pi_T s_5 & \pi_T s_6 & * \end{pmatrix}$$

- (j)Modeltest hierarchy Posada and Crandell, 1998
- Huelsenback *et. al.* 2004 considered submodels via constraints on the "relative rates" s_i

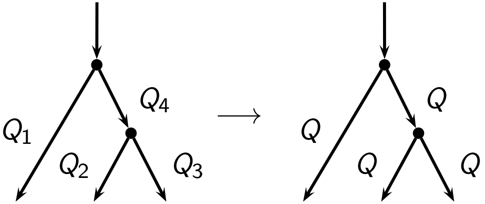
  I emailed this paper to Peter Jarvis in 2009...

What about the homogeneity assumption?

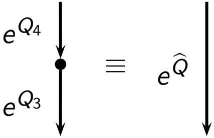
- Phylogenetic models are full of contradictory assumptions (of course!)
- Typically, substitution rates Q are assumed fixed throughout evolutionary history.
- Some modern implementations allow for differing rates on each branch.
- Leads to a problem...

What's the problem with global homogeneity?

REALITY? MODEL



Forgot 2nd taxa



$$e^{\widehat{Q}} = e^{Q_3} e^{Q_4}$$

- Is $\widehat{Q}$ in the same model?

$$\begin{aligned}\widehat{Q} &= \log\left(\exp(Q_3)\exp(Q_4)\right)\\ &= Q_3 + Q_4 + \frac{1}{2}[Q_3, Q_4] + \frac{1}{12}[Q_3,[Q_3,Q_4]] - \frac{1}{12}[Q_4,[Q_3,Q_4]] + \ldots\end{aligned}$$

- BCH formula with *commutators* $[Q_3, Q_4] := Q_3 Q_4 - Q_4 Q_3$

GTR (obviously) doesn't form a Lie algebra

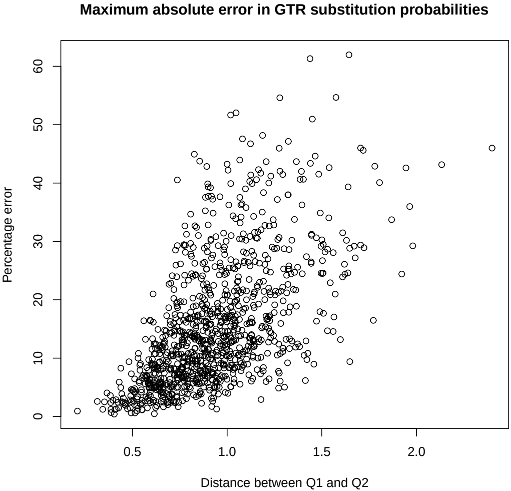
$$Q = \begin{pmatrix} * & \pi_A s_1 & \pi_A s_2 & \pi_A s_3 \\ \pi_G s_1 & * & \pi_G s_4 & \pi_G s_5 \\ \pi_C s_2 & \pi_C s_4 & * & \pi_C s_6 \\ \pi_T s_3 & \pi_T s_5 & \pi_T s_6 & * \end{pmatrix}$$

Non-linear: $q_{AG}q_{GC}q_{CA} = (\pi_A s_1)(\pi_C s_2)(\pi_G s_4) = q_{AC}q_{CG}q_{GA}$

Therefore, GTR is not multiplicatively closed.

Is the GTR model bad for molecular phylogenetics?

S *et. al. Syst. Biol.* 2012



“Almost” Lie-Markov: GTR with uniform base frequencies

- What about if $\pi_i = \frac{1}{4}$ in the GTR model?
- In this case we *do* have a linear model:

$$Q = \begin{pmatrix} * & s_1 & s_2 & s_3 \\ s_1 & * & s_4 & s_5 \\ s_2 & s_4 & * & s_6 \\ s_3 & s_5 & s_6 & * \end{pmatrix}, \qquad \text{i.e. } Q^T = Q.$$

- Since this is a linear model, via $Q^T = Q$, the first term in the BCH formula works: $(Q_1 + Q_2)^T = Q_1^T + Q_2^T$
- Commutators don't work though: $[A, B]^T = (AB)^T - (BA)^T = BA - AB = -[A, B]$
- In practice errors are not so bad up to order $\mathcal{O}(t^2)$.
- Will come back to the “dual” case $s_i = const.$ later...

Bring me a list of all Lie-Markov models. . .

1. Some (specific and general) models are already closed.
   - e.g. Kimura models, Jukes-Cantor, Felsenstein 81
   - e.g. "Group-based" and equivariant
2. What is the Lie-algebraic *closure* of a model?
   - e.g. $\overline{GTR} = GM$ and $\overline{HKY} = RY8.8$
3. Use regular representation of a finite semigroup.
   - e.g. "Group-based" and F81 (see later)
4. Constrain problem using symmetries and apply sledgehammer. $\checkmark\checkmark\checkmark$

Purine/pyrimidine symmetries

- Nucleotides can be divided into purines $\{A, G\}$ and pyrimidines $\{C, T\}$
- Purines: 2 carbon-nitrogen ring, Pyrimidines: 1 carbon-nitrogen ring

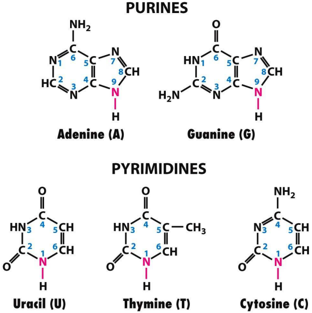


Figure 2-17
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

Models with purine/pyrimidine symmetry

- Kimura 2-parameter stationary (K2ST) 1980:

$$Q = \begin{pmatrix} * & \alpha & \beta & \beta \\ \alpha & * & \beta & \beta \\ \beta & \beta & * & \alpha \\ \beta & \beta & \alpha & * \end{pmatrix}$$

- Hasegawa, Kishino and Yano (HKY) 1985:

$$Q = \begin{pmatrix} * & \pi_A\alpha & \pi_A\beta & \pi_A\beta \\ \pi_G\alpha & * & \pi_G\beta & \pi_G\beta \\ \pi_C\beta & \pi_C\beta & * & \pi_C\alpha \\ \pi_T\beta & \pi_T\beta & \pi_T\alpha & * \end{pmatrix}$$

Purine/pyrimidine symmetries

- The mathematicians view $AG|CT = \{\{A, G\}, \{C, T\}\}$
- Symmetries: (AG) and $(AC)(GT) \in \mathfrak{S}_4$
- Generates the dihedral group $D_8 \cong C_2 \wr C_2$:

$$\{e, (AG), (CT), (AG)(CT), (AC)(GT), (AT)(GC), (ACGT), (ATGC)\}$$

Purine/pyrimidine symmetries

- The mathematicians view $AG|CT = \{\{A,G\},\{C,T\}\}$
- Symmetries: (AG) and $(AC)(GT) \in \mathfrak{S}_4$
- Generates the dihedral group $D_8 \cong C_2 \wr C_2$:

$$\{e,(AG),(CT),(AG)(CT),(AC)(GT),(AT)(GC),(ACGT),(ATGC)\}$$

Example with $\sigma = (AC)(GT)$:

$$Q = \begin{pmatrix} * & \pi_A\alpha & \pi_A\beta & \pi_A\beta \\ \pi_G\alpha & * & \pi_G\beta & \pi_G\beta \\ \pi_C\beta & \pi_C\beta & * & \pi_C\alpha \\ \pi_T\beta & \pi_T\beta & \pi_T\alpha & * \end{pmatrix} \to \begin{pmatrix} * & \pi_C\alpha & \pi_C\beta & \pi_C\beta \\ \pi_T\alpha & * & \pi_T\beta & \pi_T\beta \\ \pi_A\beta & \pi_A\beta & * & \pi_A\alpha \\ \pi_G\beta & \pi_G\beta & \pi_G\alpha & * \end{pmatrix}$$

- The labels change but this is still a HKY rate matrix!

UNIVERSITY of TASMANIA

Enter more algebra: group representation theory

- All popular models (Lie-Markov or not) have some permutation symmetries.
- e.g. GM, GTR, JC, K3ST, F81 have complete symmetry.

  K3ST:

$$Q = \begin{pmatrix} * & \alpha & \beta & \gamma \\ \alpha & * & \gamma & \beta \\ \beta & \gamma & * & \alpha \\ \gamma & \beta & \alpha & * \end{pmatrix}$$

- e.g. K2ST, HKY have purine/pyrimidine symmetry.
- Algebraic theory says (for a linear model!) we can decompose into a sum of irreducible representations of the relevant permutation group.
- e.g. F81$\cong$ $id \oplus (31)$ and K3ST$\cong$ $id \oplus (2^2)$
- In other words 4=1+3 and 3=1+2.

What the hell does $F81 \cong id \oplus (31)$ mean?

What the hell does $F81 \cong id \oplus (31)$ mean?

- F81:

$$Q = \begin{pmatrix} * & \pi_1 & \pi_1 & \pi_1 \\ \pi_2 & * & \pi_2 & \pi_2 \\ \pi_3 & \pi_3 & * & \pi_3 \\ \pi_4 & \pi_4 & \pi_4 & * \end{pmatrix} = \pi_1 R_1 + \pi_2 R_2 + \pi_3 R_3 + \pi_4 R_4$$

- The matrices $\{R_1, R_2, R_3, R_4\}$ form a basis for this model, e.g.:

$$R_1 = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

- Under permutations $\sigma \in \mathfrak{S}_4$ clearly $R_i \mapsto R_{\sigma(i)}$

  i.e. F81 forms a representation of $\mathfrak{S}_4$.
- *id* is the trivial part: $R_1 + R_2 + R_3 + R_4$ (i.e. JC model!)
- (31) is what's left over: $\{R_1 - R_2, R_1 - R_3, R_1 - R_4\}$

What was that about a sledgehammer?

- F81 is a Lie-Markov model: $[R_i, R_j] = R_i - R_j$
- We can form the analogous model with constant columns:

$$Q = \begin{pmatrix} * & \alpha_2 & \alpha_3 & \alpha_4 \\ \alpha_1 & * & \alpha_3 & \alpha_4 \\ \alpha_1 & \alpha_2 & * & \alpha_4 \\ \alpha_1 & \alpha_2 & \alpha_3 & * \end{pmatrix} = \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4$$

- Again this provides the *id* $\oplus$ (31) representation of $\mathfrak{S}_4$...

What was that about a sledgehammer?

- F81 is a Lie-Markov model: $[R_i, R_j] = R_i - R_j$
- We can form the analogous model with constant columns:

$$Q = \begin{pmatrix} * & \alpha_2 & \alpha_3 & \alpha_4 \\ \alpha_1 & * & \alpha_3 & \alpha_4 \\ \alpha_1 & \alpha_2 & * & \alpha_4 \\ \alpha_1 & \alpha_2 & \alpha_3 & * \end{pmatrix} = \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4$$

- Again this provides the $id \oplus (31)$ representation of $\mathfrak{S}_4 \ldots$

But this is not a Lie-Markov model!

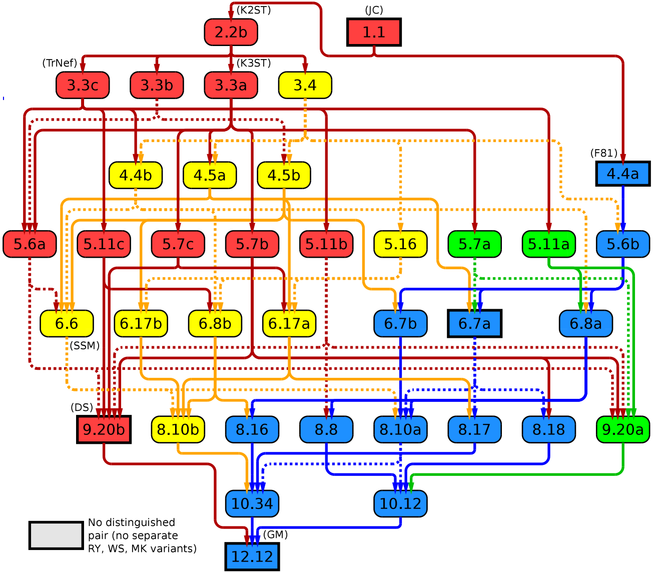
$$[C_1, C_2] = \begin{pmatrix} -3 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & -3 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} - C_2 C_1 = \begin{pmatrix} -1 & 3 & 0 & 0 \\ 3 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \end{pmatrix} \; X$$

Her All-embracing Majesty, the general Markov model (!Weyl ∼1939)

- GM$\cong$ $id \oplus 2(31) \oplus (2^2) \oplus (21^3)$
- In other words: $12 = 1 + 2 \times 3 + 2 + 3$.
- **Our big idea**: *Models with symmetries must come as direct sum of irreducible bits*

  Reduces computational complexity of "use a sledgehammer" approach just enough to solve the problem.
- i.e. $id \oplus (31) = \{R_1, R_2, R_3, R_4\}$ and $\{C_1, C_2, C_3, C_4\}$
- **Result** (S et. al. JTB 2012): The Lie subalgebras of GM with full symmetry are JC, K3ST, F81, F+K, and GM.
- **Result** (Fernandez-Sanchez *et. al.* JMB 2015): There are (roughly) 35 Lie-Markov models with purine/pyrimidine symmetry.

The Lie-Markov models with purine/pyrimidine symmetry



No distinguished pair (no separate RY, WS, MK variants)

Equilibrium base frequencies:

$\pi_A=\pi_G=\pi_C=\pi_T$ | $\pi_A=\pi_G\,;\,\pi_C=\pi_T$ | $\pi_A+\pi_G=\pi_C+\pi_T$ | $\pi_A\neq\pi_G\neq\pi_C\neq\pi_T$

“More than” Lie-Markov

- A *matrix algebra* $\mathcal{A}$ (as opposed to a Lie algebra), is a linear set of matrices closed under products: $AB \in \mathcal{A}$
- All matrix algebras form Lie algebras automatically: $[A, B] := AB - BA \in \mathcal{A}$
- The reverse is not true (see next slide for counter example).
- In our 2015 characterization of models with purine/pyrimidine symmetry each model we found actually forms a matrix algebra.
- This is *probably* because the symmetry conditions are so strong.
- So do the “equivariant” models (Draisma and Kuttler 2009).

“More than” Lie-Markov: Noether’s central dogma

- Any semi-group produces a Lie-Markov model under the *regular representation*, as follows.
- Consider the semigroup S with products $xy = x$.

  If $S = \{a_1, a_2, a_3, a_4\}$ we have, e.g.:

$$a_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

“More than” Lie-Markov: Noether’s central dogma

- Any semi-group produces a Lie-Markov model under the *regular representation*, as follows.
- Consider the semigroup S with products $xy = x$.
  If $S = \{a_1, a_2, a_3, a_4\}$ we have, e.g.:

$$a_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

- This produces a Lie-Markov model: $R_i := -\mathbf{1} + a_i$ satisfying $[R_i, R_j] = [a_i, a_j] = a_i - a_j = R_i - R_j$.

“More than” Lie-Markov: Noether’s central dogma

- Any semi-group produces a Lie-Markov model under the *regular representation*, as follows.
- Consider the semigroup S with products $xy = x$.
  If $S = \{a_1, a_2, a_3, a_4\}$ we have, e.g.:

$$a_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

- This produces a Lie-Markov model: $R_i := -\mathbf{1} + a_i$ satisfying $[R_i, R_j] = [a_i, a_j] = a_i - a_j = R_i - R_j$.
- None other than the F81 model!

$$Q = \pi_1 R_1 + \pi_2 R_2 + \pi_3 R_3 + \pi_4 R_4 = \begin{bmatrix} * & \pi_1 & \pi_1 & \pi_1 \\ \pi_2 & * & \pi_2 & \pi_2 \\ \pi_3 & \pi_3 & * & \pi_3 \\ \pi_4 & \pi_4 & \pi_4 & * \end{bmatrix}$$

“Exactly” Lie-Markov

- We know a few Lie-Markov models which *do not* form matrix algebras.
  - **i.** “Symmetric embedded” Jarvis and **S** 2012.
  - **ii.** *AustMS 2015* model:

$$L_1 = \begin{bmatrix} -3 & 0 & 0 \\ 1 & 0 & 1 \\ 2 & 0 & -1 \end{bmatrix} \quad L_2 = \begin{bmatrix} -1 & 0 & 2 \\ 1 & 0 & 1 \\ 0 & 0 & -3 \end{bmatrix}$$

- Both models satisfy $[L_1, L_2] = L_1 - L_2$, but have algebraic closures $\{L_1, L_2, X, Y, Z\}$ and $\{L_1, L_2, X\}$ respectively.
- e.g.

$$L_1^2 = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 0 & 0 \\ -6 & 0 & 0 \end{bmatrix} = -3L_1 + L_2 + X = -3L_1 + L_2 + \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 2 \\ -2 & 0 & -2 \end{bmatrix}$$

Final thoughts

- Does anyone here have any other ideas on how to proceed?
- Does this issue matter in other contexts?
- Can the Lie-Markov condition be used as a productive constraint in other contexts?
- What about time-inhomogeneous Markov chains? What is the Lie-Markov condition saying in this case?

References

Sumner JG, Fernandez-Sanchez J, Jarvis PD. 2012. Lie Markov models. Journal of Theoretical Biology

Fernandez-Sanchez J, Sumner JG, Jarvis PD, Woodhams MD. 2015. Lie Markov models with purine/pyrimidine symmetry. Journal of Mathematical Biology

Draisma J, Kuttler J. 2009. On the ideals of equivariant tree models. Mathematische Annalen

Sumner, Holland, Woodhams, Kaine, Jarvis, Fenandez-Sanchez (2012). Is GTR bad for molecular phylogenetics?. Syst. Biol.

JP Huelsenbeck, B Larget, ME Alfaro (2004). Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol.

D Posada, KA Crandall (1998). Modeltest: testing the model of DNA substitution. Bioinformatics.

Tavare S (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura RM, editor. Lectures on mathematics in the life sciences. Volume 17. Providence (RI): American Mathematical Society. p. 57-86.