Accomodating rate heterogeneity to substitution models

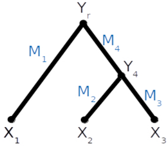
Marta Casanellas Rius

Universitat Politècnica de Catalunya
Centre de Recerca Matemàtica

ICERM, October 23rd 2024

Modeling character substitution on a tree

- At each node: random variable with k states, e.g. $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$

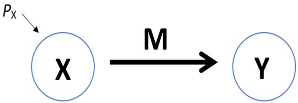


Parameters:

- distribution $\pi^r = (\pi_\mathtt{A}, \pi_\mathtt{C}, \pi_\mathtt{G}, \pi_\mathtt{T})$ at the root,
- substitution matrices: conditional probabilities of substitution,

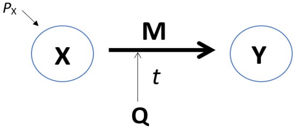
$$M_i = \begin{array}{c c} & \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{array}{c} \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} & \begin{pmatrix} P(\mathtt{A}|\mathtt{A}) & P(\mathtt{C}|\mathtt{A}) & P(\mathtt{G}|\mathtt{A}) & P(\mathtt{T}|\mathtt{A}) \\ P(\mathtt{A}|\mathtt{C}) & P(\mathtt{C}|\mathtt{C}) & P(\mathtt{G}|\mathtt{C}) & P(\mathtt{T}|\mathtt{C}) \\ P(\mathtt{A}|\mathtt{G}) & P(\mathtt{C}|\mathtt{G}) & P(\mathtt{G}|\mathtt{G}) & P(\mathtt{T}|\mathtt{G}) \\ P(\mathtt{A}|\mathtt{T}) & P(\mathtt{C}|\mathtt{T}) & P(\mathtt{G}|\mathtt{T}) & P(\mathtt{T}|\mathtt{T}) \end{pmatrix} \end{array}$$

Substitution process on a single edge



- no assumption on the underlying process
- M: Markov matrix
- $k \times k$ non-negative matrix, sum of rows equal to one
- for $k = 4$, 12 free parameters
- for $k = 20$, 380 free parameters

Continuous-time process on an edge



- **(local homogeneous) continuous time**: instantaneous rates of substitution have the same shape along the process,
- collected in a **rate matrix** Q

$$Q=\begin{pmatrix} \bullet & q_{\mathtt{A},\mathtt{C}} & q_{\mathtt{A},\mathtt{G}} & q_{\mathtt{A},\mathtt{T}} \\ q_{\mathtt{C},\mathtt{A}} & \bullet & q_{\mathtt{C},\mathtt{G}} & q_{\mathtt{C},\mathtt{T}} \\ q_{\mathtt{G},\mathtt{A}} & q_{\mathtt{G},\mathtt{C}} & \bullet & q_{\mathtt{G},\mathtt{T}} \\ q_{\mathtt{T},\mathtt{A}} & q_{\mathtt{T},\mathtt{C}} & q_{\mathtt{T},\mathtt{G}} & \bullet \end{pmatrix}$$

- rows sum to 0
- $q_{i,j} \geq 0$, $i \neq j$.

- Transition matrix $M(t)$ satisfies $M'(t) = M(t)Q$, $M(0) = Id$.

$$M = \exp(tQ)$$

Continuous versus general Markov on a single edge

- Question (The embedding problem, Elfving, 1937):
  When is a Markov matrix M the exponential of a rate matrix?

  $$M = \exp(tQ)$$

  These are called **embeddable** matrices
- Only solved for 2×2 and 3×3 matrices until ’23
- We provide al algorithm[1] to test embeddability. For 4×4 matrices this gives:
- $< 1\%$ of Markov matrices are of this type
- Restricting to Diagonal Largest in Column: $< 4\%$
- Restricting to Diagonally Dominant: $\sim 12\%$

[1] C–Fernández-Sánchez–Roca-Lacostena, The embedding problem for Markov matrices, Publicacions Matemàtiques 2023

Continuous versus general Markov on a single edge

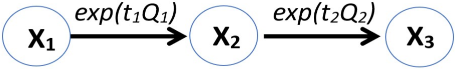
- Question (The embedding problem, Elfving, 1937):
  When is a Markov matrix M the exponential of a rate matrix?

$$M = \exp(tQ)$$

  These are called **embeddable** matrices
- Only solved for 2×2 and 3×3 matrices until '23
- We provide al algorithm[1] to test embeddability. For 4×4 matrices this gives:
- $< 1\%$ of Markov matrices are of this type
- Restricting to Diagonal Largest in Column: $< 4\%$
- Restricting to Diagonally Dominant: $\sim 12\%$

[1] C–Fernández-Sánchez–Roca-Lacostena, The embedding problem for Markov matrices, Publicacions Matemàtiques 2023

Local to Global homogeneity



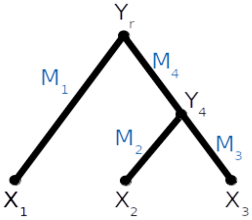
- Problem: concatenating time-continuous processes does not give an embeddable matrix

$$exp(t_1 Q_1) exp(t_2 Q_2) \neq exp(tQ)$$

- Most models are NOT multiplicatively closed [2]
- Global homogeneity of rates in continuous-time models: Same rate matrix Q at all edges (multiplicatively closed)

[2] Sumner et al., Sys Bio 2012

On a tree: Continuous-time vs. general Markov



By considering a general Markov (GM) process we allow

- local heterogeneity: change of rates along an edge
- global heterogeneity: different rates at different lineages

Parameters of GM

- Amount of parameters: 12 (or 380) times the number of edges + distribution at root.
- A maximum-likelihood for topology reconstruction approach is impractical

Alternative approaches: SVD, based on phylogenetic invariants theory

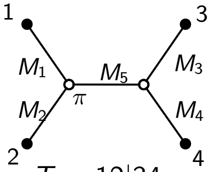
Flattening and SVD

16×16 matrix obtained by *flattening*
$p = (p_{\texttt{AAAA}}, p_{\texttt{AAAC}}, \ldots, p_{\texttt{TTTT}})$

$$\mathit{flat}_{12|34}(p) = \begin{array}{cc} & \textit{states at } 3,4 \\ \begin{array}{c}\textit{states}\\ \textit{at}\\ 1,2\end{array} & \begin{pmatrix} p_{\texttt{AAAA}} & p_{\texttt{AAAC}} & p_{\texttt{AAAG}} & \cdots & p_{\texttt{AATT}} \\ p_{\texttt{ACAA}} & p_{\texttt{ACAC}} & p_{\texttt{ACAG}} & \cdots & p_{\texttt{ACTT}} \\ p_{\texttt{AGAA}} & p_{\texttt{AGAC}} & p_{\texttt{AGAG}} & \cdots & p_{\texttt{AGTT}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{\texttt{TTAA}} & p_{\texttt{TTAC}} & p_{\texttt{TTAG}} & \cdots & p_{\texttt{TTTT}} \end{pmatrix} \end{array}$$

- (Allman-Rhodes'08) If $p = p^T \Rightarrow \operatorname{rank}(\mathit{flat}_{12|34}\, p) \leq 4$
- for $T = 13|24$, $14|23$, rank 16 (in general)

Flattening and SVD



$T = 12|34$

16×16 matrix obtained by *flattening*
$p = (p_{\texttt{AAAA}}, p_{\texttt{AAAC}}, \ldots, p_{\texttt{TTTT}})$

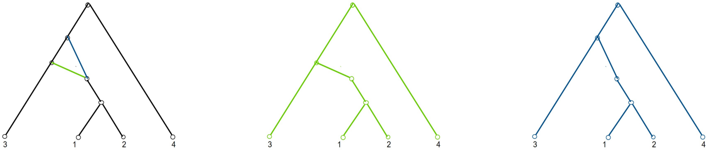
$$flat_{12|34}(p) = \begin{array}{cc} & \textit{states at } 3,4 \\ \begin{array}{c}\textit{states}\\ \textit{at}\\ 1,2\end{array} & \begin{pmatrix} p_{\texttt{AAAA}} & p_{\texttt{AAAC}} & p_{\texttt{AAAG}} & \cdots & p_{\texttt{AATT}} \\ p_{\texttt{ACAA}} & p_{\texttt{ACAC}} & p_{\texttt{ACAG}} & \cdots & p_{\texttt{ACTT}} \\ p_{\texttt{AGAA}} & p_{\texttt{AGAC}} & p_{\texttt{AGAG}} & \cdots & p_{\texttt{AGTT}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{\texttt{TTAA}} & p_{\texttt{TTAC}} & p_{\texttt{TTAG}} & \cdots & p_{\texttt{TTTT}} \end{pmatrix} \end{array}$$

- (Allman-Rhodes'08) If $p = p^T \Rightarrow \text{rank}(flat_{12|34}\, p) \leq 4$
- for $T = 13|24$, $14|23$, rank 16 (in general)

SVD approach

- Valid for GM model, any number of states k (rank $\leq k$)
- Singular value decomposition (SVD): to test how far is a matrix from rank k
- This has been used in: Erik+2, Splitscores, SVDQuartets, SAQ and ASAQ
- Quartet-based

Works for some phylogenetic networks



$$p \quad = \quad \delta p^{T_1} \quad + \quad (1-\delta) p^{T_2}$$

Theorem (C–Fernández-Sánchez'21)

$rank(flat_{12|34}\, p) \leq 4$ if p is a distribution on this network.

Networks on n leaves: if the network has a tree clade T_A, $flat_{A|B}(p)$ has rank ≤ 4.

Flattening for more restrictive models

- GM is probably too complex for what is used nowadays with amino acids (empirical models)
- Other models that allow heterogeneity of rates:

Kimura 3-parameter model (K81)

- π uniform distribution: $\pi_{\mathtt{A}} = \pi_{\mathtt{C}} = \pi_{\mathtt{G}} = \pi_{\mathtt{T}} = 0.25$
- M^e: 3 free parameters per edge

$$M^e = \begin{array}{c c} & \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{array}{c} A \\ C \\ G \\ T \end{array} & \begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix} \end{array}, \quad a + b + c + d = 1$$

- K80, JC69 submodels

Fourier Coordinates for Group-based models

Hadamard transform (90's Erdös, Székely, Hendy, Penny, Steel, Evans, Speed):

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Linear change of coordinates: $\bar{p} = (H \otimes \cdots \otimes H)^{-1} p$

Equivalently: Fourier basis

$$u^1 = \tfrac{1}{4}(1,1,1,1)^t \qquad u^2 = \tfrac{1}{4}(1,1,-1,-1)^t$$
$$u^3 = \tfrac{1}{4}(1,-1,1,-1)^t \qquad u^4 = \tfrac{1}{4}(1,-1,-1,1)^t$$

- All K81 matrices diagonalize in this basis.
- If $p \in \mathbb{R}^4 \otimes \cdots \otimes \mathbb{R}^4 \to \bar{p} = (\bar{p}_{1\ldots1}, \ldots, \bar{p}_{4\ldots4})$: coordinates in basis

$$u^1 \otimes \cdots \otimes u^1, \quad u^1 \otimes \cdots \otimes u^2, \quad \ldots, \quad u^4 \otimes \cdots \otimes u^4$$

Flattening for K81

If $\overline{p}$ are the Fourier coordinates of p, reordering rows and columns, $flat_{A|B}(\bar{p})$ is block-diagonal

$$flat_{12|34}(\overline{p}) = \begin{pmatrix} B_1 & 0 & 0 & 0 \\ 0 & B_2 & 0 & 0 \\ 0 & 0 & B_3 & 0 \\ 0 & 0 & 0 & B_4 \end{pmatrix}$$

$$4 \times 4 \text{ blocks } B_1, B_2, B_3, B_4$$

Theorem (Draisma-Kuttler'09, C–Fernández-Sánchez'11)

If $T = 12|34$ and p is a distribution on T under K81 model, then

$$\mathbf{rk}(B_1, B_2, B_3, B_4) \leq (1, 1, 1, 1).$$

Flattening for K81

If $\overline{p}$ are the Fourier coordinates of p, reordering rows and columns, $flat_{A|B}(\bar{p})$ is block-diagonal

$$flat_{12|34}(\overline{p}) = \begin{pmatrix} B_1 & 0 & 0 & 0 \\ 0 & B_2 & 0 & 0 \\ 0 & 0 & B_3 & 0 \\ 0 & 0 & 0 & B_4 \end{pmatrix}$$

$$4 \times 4 \text{ blocks } B_1, B_2, B_3, B_4$$

Theorem (Draisma-Kuttler'09, C–Fernández-Sánchez'11)

If $T = 12|34$ and p is a distribution on T under K81 model, then

$$\mathbf{rk}(B_1, B_2, B_3, B_4) \leq (1, 1, 1, 1).$$

Flattening for JC69

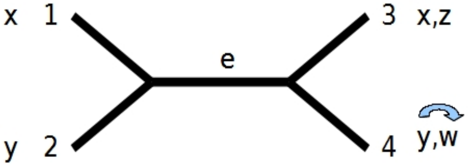
In a certain basis $flat_{12|34}\bar{p}$ can be reduced to

$$\left(\begin{array}{c||c|c|c} \begin{matrix} * & * \\ * & * \end{matrix} & 0 & 0 & 0 \\ \hline\hline 0 & * & 0 & 0 \\ \hline 0 & 0 & \begin{matrix} * & * & * \\ * & * & * \\ * & * & * \end{matrix} & 0 \\ \hline 0 & 0 & 0 & * \end{array}\right)$$

Theorem (C–Fernández-Sánchez'11) $\mathbf{rk} \leq (1,0,0,1,0)$.
Consequence: linear equations equivalent to Lake's invariants,

$$\bar{p}_{2222} = \bar{p}_{2244} \qquad \bar{p}_{2424} = \bar{p}_{2442}$$

Lake's linear invariants (1987)



For the JC69 and K81 model on the tree $12|34$ the following are linear topology invariants:

$$H_1: \quad p_{xyxy} + p_{xyzw} = p_{xyzy} + p_{xyxw}$$

$$H_2: \quad p_{xyyx} + p_{xywz} = p_{xyyz} + p_{xywx}$$

for any x, y, z, w in $\{A, C, G, T\}$.

- H_1 is NOT a phylogenetic invariant for $13|24$ and H_2 is NOT an invariant for $14|23$.

Looking for in-between models

Looking for models such that

- can be defined on any number of states
- do not assume continuous-time, allow heterogeneous rates

But

- GM might be too general
- Group-based models too restrictive (not for any number of states, stationary distribution is uniform)

In between: time-reversible models

Stationary and time-reversible models

k states

- Markov matrices have a stationary distribution π: $\pi^t M = \pi^t$
- A Markov process $X \xrightarrow{M} Y$ is time-reversible if $Pr(X=i, Y=j) = Pr(X=j, Y=i)$ at equilibrium:

$$\pi_i M_{i,j} = \pi_j M_{j,i}.$$

- Fix a distribution $\pi = (\pi_1, \ldots, \pi_k)$.
  A Markov matrix is π-time-reversible if $D_\pi M = M^t D_\pi$
  (and then π is its stationary distribution)

Stationary and time-reversible models

k states

- Markov matrices have a stationary distribution π: $\pi^t M = \pi^t$
- A Markov process $X \xrightarrow{M} Y$ is time-reversible if $Pr(X = i, Y = j) = Pr(X = j, Y = i)$ at equilibrium:

$$\pi_i M_{i,j} = \pi_j M_{j,i}.$$

- Fix a distribution $\pi = (\pi_1, \ldots, \pi_k)$.
  A Markov matrix is π-time-reversible if $D_\pi M = M^t D_\pi$
  (and then π is its stationary distribution)

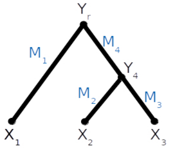
Algebraic time-reversible processes

A Markov process on a tree is

- stationary if **all** transition matrices have the same stationary distribution π
- π-time-reversible if **all** transition matrices are π-time-reversible.
- Time-reversible process $\Rightarrow \pi =$ distribution at root

Definition

Algebraic time-reversible[3] (ATR) process: All transition matrices are π-time-reversible and **commute**.



(if $M_i = e^{t_i Q} \Rightarrow$ commute)

[3] Allman-Rhodes, J. Symbolic Comput. 2006

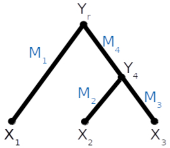
Algebraic time-reversible processes

A Markov process on a tree is

- stationary if **all** transition matrices have the same stationary distribution π
- π-time-reversible if **all** transition matrices are π-time-reversible.
- Time-reversible process $\Rightarrow \pi =$ distribution at root

Definition

Algebraic time-reversible[3] (ATR) process: All transition matrices are π-time-reversible and **commute**.



(if $M_i = e^{t_i Q} \Rightarrow$ commute)

[3] Allman-Rhodes, J. Symbolic Comput. 2006

Examples of ATR models

Ex: homogeneous GTR
Ex: group-based models
Ex: Tamura-Nei model (TN93)

$$
\begin{array}{c}
\begin{array}{cccc} \quad\mathtt{A} & \quad\mathtt{G} & \quad\mathtt{C} & \quad\mathtt{T} \end{array} \\
M = \begin{pmatrix}
*_1 & \pi_2 c & \pi_3 b & \pi_4 b \\
\pi_1 c & *_2 & \pi_3 b & \pi_4 b \\
\pi_1 b & \pi_2 b & *_3 & \pi_4 d \\
\pi_1 b & \pi_2 b & \pi_3 d & *_4
\end{pmatrix}
\end{array}
$$

- Submodels: HKY85, F81

Equal-Input model (EI)

- π : distribution on k states (stationary distribution)
- Equal Input model: for each edge e of T conditional probabilities satisfy:

$$Prob(y|x) = \pi_y \cdot a_e, \text{for some } a_e \in [0,1]$$

$$M = \begin{pmatrix} * & \pi_2 a & \dots & \pi_k a \\ \pi_1 a & * & \dots & \pi_k a \\ & \vdots & \vdots & \\ \pi_1 a & \pi_2 a & \dots & \pi_k a \\ \pi_1 a & \pi_2 a & \dots & * \end{pmatrix}$$

- For $k = 4$, this is F81 model
- If π is uniform, it's the Fully symmetric model (JC69 for $k = 4$, CFN for $k = 2$)

Equal-Input model (EI)

- π : distribution on k states (stationary distribution)
- Equal Input model: for each edge e of T conditional probabilities satisfy:

$$Prob(y|x) = \pi_y \cdot a_e, \text{for some } a_e \in [0,1]$$

$$M = \begin{pmatrix} * & \pi_2 a & \dots & \pi_k a \\ \pi_1 a & * & \dots & \pi_k a \\ & \vdots & \vdots & \\ \pi_1 a & \pi_2 a & \dots & \pi_k a \\ \pi_1 a & \pi_2 a & \dots & * \end{pmatrix}$$

- For $k = 4$, this is F81 model
- If π is uniform, it's the Fully symmetric model (JC69 for $k = 4$, CFN for $k = 2$)

Towards a non-uniform π (+R. Homs, A. Torres)

- ATR: substitution matrices
  commute $\leftrightarrow$ simultaneously diagonalizable
- Fourier basis $\rightarrow$ **orthogonal** basis of eigenvectors
- K81 matrices are symmetric ($\Rightarrow$ Spectral theorem)
- For K81, π is uniform and does not play any rol

Goal: Generalize these tools to

- non-uniform π
- any number k of states

π-time-reversible model $\Rightarrow$ π can be estimated from data $\Rightarrow$ fixed

Definition (π-inner product)

$$\langle u, v\rangle_\pi := \sum_i \frac{1}{\pi_i} u_i v_i = u^t D_\pi^{-1} v$$

Towards a non-uniform π (+R. Homs, A. Torres)

- ATR: substitution matrices

  commute $\leftrightarrow$ simultaneously diagonalizable
- Fourier basis $\rightarrow$ **orthogonal** basis of eigenvectors
- K81 matrices are symmetric ($\Rightarrow$ Spectral theorem)
- For K81, π is uniform and does not play any rol

Goal: Generalize these tools to

- non-uniform π
- any number k of states

π-time-reversible model $\Rightarrow$ π can be estimated from data $\Rightarrow$ fixed

Definition (π-inner product)

$$\langle u, v\rangle_\pi := \sum_i \frac{1}{\pi_i} u_i v_i = u^t D_\pi^{-1} v$$

Lemma: M is π time-reversible $\Leftrightarrow$ M^t is self-adjoint for $\langle\, ,\rangle_\pi$

We can use Spectral Theorem

B-time-reversible

- ATR $\Rightarrow$ simultaneously diagonalizable & exists π-orthogonal eigenbasis for M^t
- Let $B = \{u^1 = \pi, \ldots, u^k\}$ be a π-orthogonal basis in $\mathbb{R}^k$,
- M has B as left-eigenbasis $\Rightarrow$ M is π-time-reversible

Definition
B-time-reversible model on a phylogenetic tree T: all transition matrices have B as left-eigenbasis

- B =Fourier $\Rightarrow$ K81, K80 and JC69 are B-time-reversible

B-time-reversible

- ATR $\Rightarrow$ simultaneously diagonalizable & exists π-orthogonal eigenbasis for M^t
- Let $B = \{u^1 = \pi, \ldots, u^k\}$ be a π-orthogonal basis in $\mathbb{R}^k$,
- M has B as left-eigenbasis $\Rightarrow$ M is π-time-reversible

Definition

B-time-reversible model on a phylogenetic tree T: all transition matrices have B as left-eigenbasis

- B =Fourier $\Rightarrow$ K81, K80 and JC69 are B-time-reversible

Example: TN93

$$M = \begin{pmatrix} *_1 & \pi_2 c & \pi_3 b & \pi_4 b \\ \pi_1 c & *_2 & \pi_3 b & \pi_4 b \\ \pi_1 b & \pi_2 b & *_3 & \pi_4 d \\ \pi_1 b & \pi_2 b & \pi_3 d & *_4 \end{pmatrix}$$

$$B = \left\{ \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix}, \begin{pmatrix} \pi_1\pi_{34} \\ \pi_2\pi_{34} \\ -\pi_3\pi_{12} \\ -\pi_4\pi_{12} \end{pmatrix}, \frac{1}{\pi_{34}} \begin{pmatrix} 0 \\ 0 \\ \pi_3\pi_4 \\ -\pi_3\pi_4 \end{pmatrix}, \frac{1}{\pi_{12}} \begin{pmatrix} \pi_1\pi_2 \\ -\pi_1\pi_2 \\ 0 \\ 0 \end{pmatrix} \right\}$$

- B is a left-eigenbasis for $N \Leftrightarrow N$ is a TN93 matrix
- Submodels: HKY85, F81 are B-time reversible

New coordinates and reparameterization

π fixed, Markov process on T parametrized as:

$$\begin{array}{rcl} Parameters & \xrightarrow{\psi_T} & \bigotimes^n \mathbb{R}^k \\ (M^e)_{e\in E(T)} & \mapsto & p^T = (p^T_{1\dots 1}, \dots, p^T_{k\dots k}) \end{array}$$

Basis in $\mathbb{R}^k \otimes \overset{n)}{\dots} \otimes \mathbb{R}^k$:

$$B^n = \{u^{i_1} \otimes u^{i_2} \cdots \otimes u^{i_n} \mid i_j \in [k]\},$$

- $\bar{p}$: coordinates of p in this basis
- parameters: eigenvalues $\Lambda^e = (\lambda^e_1, \dots, \lambda^e_k)$ of transition matrices M^e
- Reparameterization of Markov process on T:

$$\begin{array}{rcl} \prod_{e\in E(T)} \mathbb{R}^k & \xrightarrow{\varphi_T} & \bigotimes^n \mathbb{R}^k \\ (\Lambda^e)_{e\in E(T)} & \mapsto & \bar{p}^T = (\bar{p}^T_{1\dots 1}, \dots, \bar{p}^T_{k\dots k}) \end{array}$$

New coordinates and reparameterization

π fixed, Markov process on T parametrized as:

$$\begin{array}{ccc} Parameters & \xrightarrow{\psi_T} & \bigotimes^n \mathbb{R}^k \\ (M^e)_{e\in E(T)} & \mapsto & p^T = (p^T_{1\dots 1}, \dots, p^T_{k\dots k}) \end{array}$$

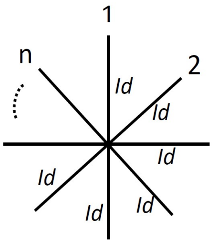
Basis in $\mathbb{R}^k \otimes \overset{n)}{\dots} \otimes \mathbb{R}^k$:

$$B^n = \{u^{i_1} \otimes u^{i_2} \cdots \otimes u^{i_n} \mid i_j \in [k]\},$$

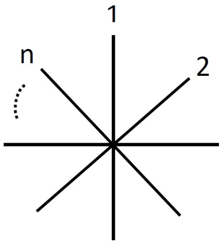
- $\bar{p}$: coordinates of p in this basis
- parameters: eigenvalues $\Lambda^e = (\lambda^e_1, \dots, \lambda^e_k)$ of transition matrices M^e
- Reparameterization of Markov process on T:

$$\begin{array}{ccc} \prod_{e\in E(T)} \mathbb{R}^k & \xrightarrow{\varphi_T} & \bigotimes^n \mathbb{R}^k \\ (\Lambda^e)_{e\in E(T)} & \mapsto & \bar{p}^T = (\bar{p}^T_{1\dots 1}, \dots, \bar{p}^T_{k\dots k}) \end{array}$$

Star trees



p^0



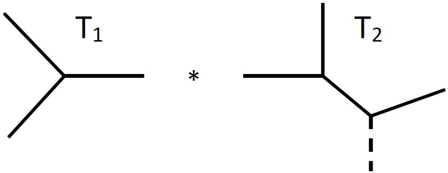
p

$$\overline{p^0} = \varphi_T(\{Id, Id, \dots, Id\}) \Rightarrow \bar{p} = (\Lambda^1 \otimes \cdots \otimes \Lambda^n)\overline{p^0}$$

Lemma

Star trees evolving under a B-time-reversible model have a monomial parameterization in these coordinates

Glueing trees



Theorem (C-Homs-Torres)

$$\bar{p}^{T}_{i_1\ldots i_n} = \sum_{j\in\Sigma} \langle u^j, u^j \rangle_\pi \, \bar{p}^{T_1}_{i_1\ldots i_m\, j} \, \bar{p}^{T_2}_{j\, i_{m+1}\ldots i_n}$$

Corollary

For group-based models, we recover Evans-Speed theorem. For other ATR models, we obtain a new framework to get phylogenetic invariants.

$flat_{12|34}$ for TN93

$(1,1)$	$(1,4)$	$(4,1)$	$(2,4)$	$(4,2)$	$(4,4)$	$(2,2)$	$(1,2)$	$(2,1)$	$(3,3)$	$(1,3)$	$(3,1)$	$(2,3)$	$(3,2)$	$(3,4)$	$(4,3$
$\bar{p}_{1111}$	0	0	0	0	$\bar{p}_{1144}$	$\bar{p}_{1122}$	0	0	$\bar{p}_{1133}$	0	0	0	0	0	0
0	$\bar{p}_{1414}$	$\bar{p}_{1441}$	$\bar{p}_{1424}$	$\bar{p}_{1442}$	$\bar{p}_{1444}$	0	0	0	0	0	0	0	0	0	0
0	$\bar{p}_{4114}$	$\bar{p}_{4141}$	$\bar{p}_{4124}$	$\bar{p}_{4142}$	$\bar{p}_{4144}$	0	0	0	0	0	0	0	0	0	0
0	$\bar{p}_{2414}$	$\bar{p}_{2441}$	$\bar{p}_{2424}$	$\bar{p}_{2442}$	$\bar{p}_{2444}$	0	0	0	0	0	0	0	0	0	0
0	$\bar{p}_{4214}$	$\bar{p}_{4241}$	$\bar{p}_{4224}$	$\bar{p}_{4242}$	$\bar{p}_{4244}$	0	0	0	0	0	0	0	0	0	0
$\bar{p}_{4411}$	$\bar{p}_{4414}$	$\bar{p}_{4441}$	$\bar{p}_{4424}$	$\bar{p}_{4442}$	$\bar{p}_{4444}$	$\bar{p}_{4422}$	$\bar{p}_{4412}$	$\bar{p}_{4421}$	$\bar{p}_{4433}$	0	0	0	0	0	0
$\bar{p}_{2211}$	0	0	0	0	$\bar{p}_{2244}$	$\bar{p}_{2222}$	$\bar{p}_{2212}$	$\bar{p}_{2221}$	$\bar{p}_{2233}$	0	0	0	0	0	0
0	0	0	0	0	$\bar{p}_{1244}$	$\bar{p}_{1222}$	$\bar{p}_{1212}$	$\bar{p}_{1221}$	$\bar{p}_{1233}$	0	0	0	0	0	0
0	0	0	0	0	$\bar{p}_{2144}$	$\bar{p}_{2122}$	$\bar{p}_{2112}$	$\bar{p}_{2121}$	$\bar{p}_{2133}$	0	0	0	0	0	0
$\bar{p}_{3311}$	0	0	0	0	$\bar{p}_{3344}$	$\bar{p}_{3322}$	$\bar{p}_{3312}$	$\bar{p}_{3321}$	$\bar{p}_{3333}$	$\bar{p}_{3313}$	$\bar{p}_{3331}$	$\bar{p}_{3323}$	$\bar{p}_{3332}$	0	0
0	0	0	0	0	0	0	0	0	$\bar{p}_{1333}$	$\bar{p}_{1313}$	$\bar{p}_{1331}$	$\bar{p}_{1323}$	$\bar{p}_{1332}$	0	0
0	0	0	0	0	0	0	0	0	$\bar{p}_{3133}$	$\bar{p}_{3113}$	$\bar{p}_{3131}$	$\bar{p}_{3123}$	$\bar{p}_{3132}$	0	0
0	0	0	0	0	0	0	0	0	$\bar{p}_{2333}$	$\bar{p}_{2313}$	$\bar{p}_{2331}$	$\bar{p}_{2323}$	$\bar{p}_{2332}$	0	0
0	0	0	0	0	0	0	0	0	$\bar{p}_{3233}$	$\bar{p}_{3213}$	$\bar{p}_{3231}$	$\bar{p}_{3223}$	$\bar{p}_{3232}$	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Rank 4, but blocks of rk 1, one of rank 2, some of rank 3, and some of rank 0

$flat_{12|34}$ for TN93

(1, 1)	(1, 4)	(4, 1)	(2, 4)	(4, 2)	(4, 4)	(2, 2)	(1, 2)	(2, 1)	(3, 3)	(1, 3)	(3, 1)	(2, 3)	(3, 2)	(3, 4)	(4, 3
$\bar{p}_{1111}$	0	0	0	0	$\bar{p}_{1144}$	$\bar{p}_{1122}$	0	0	$\bar{p}_{1133}$	0	0	0	0	0	0
0	$\bar{p}_{1414}$	$\bar{p}_{1441}$	$\bar{p}_{1424}$	$\bar{p}_{1442}$	$\bar{p}_{1444}$	0	0	0	0	0	0	0	0	0	0
0	$\bar{p}_{4114}$	$\bar{p}_{4141}$	$\bar{p}_{4124}$	$\bar{p}_{4142}$	$\bar{p}_{4144}$	0	0	0	0	0	0	0	0	0	0
0	$\bar{p}_{2414}$	$\bar{p}_{2441}$	$\bar{p}_{2424}$	$\bar{p}_{2442}$	$\bar{p}_{2444}$	0	0	0	0	0	0	0	0	0	0
0	$\bar{p}_{4214}$	$\bar{p}_{4241}$	$\bar{p}_{4224}$	$\bar{p}_{4242}$	$\bar{p}_{4244}$	0	0	0	0	0	0	0	0	0	0
$\bar{p}_{4411}$	$\bar{p}_{4414}$	$\bar{p}_{4441}$	$\bar{p}_{4424}$	$\bar{p}_{4442}$	$\bar{p}_{4444}$	$\bar{p}_{4422}$	$\bar{p}_{4412}$	$\bar{p}_{4421}$	$\bar{p}_{4433}$	0	0	0	0	0	0
$\bar{p}_{2211}$	0	0	0	0	$\bar{p}_{2244}$	$\bar{p}_{2222}$	$\bar{p}_{2212}$	$\bar{p}_{2221}$	$\bar{p}_{2233}$	0	0	0	0	0	0
0	0	0	0	0	$\bar{p}_{1244}$	$\bar{p}_{1222}$	$\bar{p}_{1212}$	$\bar{p}_{1221}$	$\bar{p}_{1233}$	0	0	0	0	0	0
0	0	0	0	0	$\bar{p}_{2144}$	$\bar{p}_{2122}$	$\bar{p}_{2112}$	$\bar{p}_{2121}$	$\bar{p}_{2133}$	0	0	0	0	0	0
$\bar{p}_{3311}$	0	0	0	0	$\bar{p}_{3344}$	$\bar{p}_{3322}$	$\bar{p}_{3312}$	$\bar{p}_{3321}$	$\bar{p}_{3333}$	$\bar{p}_{3313}$	$\bar{p}_{3331}$	$\bar{p}_{3323}$	$\bar{p}_{3332}$	0	0
0	0	0	0	0	0	0	0	0	$\bar{p}_{1333}$	$\bar{p}_{1313}$	$\bar{p}_{1331}$	$\bar{p}_{1323}$	$\bar{p}_{1332}$	0	0
0	0	0	0	0	0	0	0	0	$\bar{p}_{3133}$	$\bar{p}_{3113}$	$\bar{p}_{3131}$	$\bar{p}_{3123}$	$\bar{p}_{3132}$	0	0
0	0	0	0	0	0	0	0	0	$\bar{p}_{2333}$	$\bar{p}_{2313}$	$\bar{p}_{2331}$	$\bar{p}_{2323}$	$\bar{p}_{2332}$	0	0
0	0	0	0	0	0	0	0	0	$\bar{p}_{3233}$	$\bar{p}_{3213}$	$\bar{p}_{3231}$	$\bar{p}_{3223}$	$\bar{p}_{3232}$	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Rank 4, but blocks of rk 1, one of rank 2, some of rank 3, and some of rank 0

$flat_{12|34}$ for TN93

$(1,1)$	$(1,4)$	$(4,1)$	$(2,4)$	$(4,2)$	$(4,4)$	$(2,2)$	$(1,2)$	$(2,1)$	$(3,3)$	$(1,3)$	$(3,1)$	$(2,3)$	$(3,2)$	$(3,4)$	$(4,3$
$\bar{p}_{1111}$	0	0	0	0	$\bar{p}_{1144}$	$\bar{p}_{1122}$	0	0	$\bar{p}_{1133}$	0	0	0	0	0	0
0	$\bar{p}_{1414}$	$\bar{p}_{1441}$	$\bar{p}_{1424}$	$\bar{p}_{1442}$	$\bar{p}_{1444}$	0	0	0	0	0	0	0	0	0	0
0	$\bar{p}_{4114}$	$\bar{p}_{4141}$	$\bar{p}_{4124}$	$\bar{p}_{4142}$	$\bar{p}_{4144}$	0	0	0	0	0	0	0	0	0	0
0	$\bar{p}_{2414}$	$\bar{p}_{2441}$	$\bar{p}_{2424}$	$\bar{p}_{2442}$	$\bar{p}_{2444}$	0	0	0	0	0	0	0	0	0	0
0	$\bar{p}_{4214}$	$\bar{p}_{4241}$	$\bar{p}_{4224}$	$\bar{p}_{4242}$	$\bar{p}_{4244}$	0	0	0	0	0	0	0	0	0	0
$\bar{p}_{4411}$	$\bar{p}_{4414}$	$\bar{p}_{4441}$	$\bar{p}_{4424}$	$\bar{p}_{4442}$	$\bar{p}_{4444}$	$\bar{p}_{4422}$	$\bar{p}_{4412}$	$\bar{p}_{4421}$	$\bar{p}_{4433}$	0	0	0	0	0	0
$\bar{p}_{2211}$	0	0	0	0	$\bar{p}_{2244}$	$\bar{p}_{2222}$	$\bar{p}_{2212}$	$\bar{p}_{2221}$	$\bar{p}_{2233}$	0	0	0	0	0	0
0	0	0	0	0	$\bar{p}_{1244}$	$\bar{p}_{1222}$	$\bar{p}_{1212}$	$\bar{p}_{1221}$	$\bar{p}_{1233}$	0	0	0	0	0	0
0	0	0	0	0	$\bar{p}_{2144}$	$\bar{p}_{2122}$	$\bar{p}_{2112}$	$\bar{p}_{2121}$	$\bar{p}_{2133}$	0	0	0	0	0	0
$\bar{p}_{3311}$	0	0	0	0	$\bar{p}_{3344}$	$\bar{p}_{3322}$	$\bar{p}_{3312}$	$\bar{p}_{3321}$	$\bar{p}_{3333}$	$\bar{p}_{3313}$	$\bar{p}_{3331}$	$\bar{p}_{3323}$	$\bar{p}_{3332}$	0	0
0	0	0	0	0	0	0	0	0	$\bar{p}_{1333}$	$\bar{p}_{1313}$	$\bar{p}_{1331}$	$\bar{p}_{1323}$	$\bar{p}_{1332}$	0	0
0	0	0	0	0	0	0	0	0	$\bar{p}_{3133}$	$\bar{p}_{3113}$	$\bar{p}_{3131}$	$\bar{p}_{3123}$	$\bar{p}_{3132}$	0	0
0	0	0	0	0	0	0	0	0	$\bar{p}_{2333}$	$\bar{p}_{2313}$	$\bar{p}_{2331}$	$\bar{p}_{2323}$	$\bar{p}_{2332}$	0	0
0	0	0	0	0	0	0	0	0	$\bar{p}_{3233}$	$\bar{p}_{3213}$	$\bar{p}_{3231}$	$\bar{p}_{3223}$	$\bar{p}_{3232}$	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	**0**
0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	**0**

Rank 4, but blocks of rk 1, one of rank 2, some of rank 3, and some **rank 0**

Linear topology invariants for TN93

For $T = 12|34$, the following equalities hold:

$$\bar{p}_{3434} = 0, \bar{p}_{3443} = 0, \bar{p}_{4343} = 0, \bar{p}_{4334} = 0$$

- Linear invariants $\Rightarrow$ valid on mixtures on the same tree.
- $\bar{p}_{3434} = 0$, $\bar{p}_{4343} = 0$ hold on $T = 12|34$ and $T = 14|23$ but not on $13|24 \Rightarrow$ valid for identifying mixtures on pairs of trees.

These are generalized Lake's invariants.

Open Questions

- EI model 4 states (F81): rank of blocks of flattening?
- EI model, any number of states: change of coordinates (already working on this with G. Dilaver, J. Garbett, R. Homs, A. Korchmaros, N. Paul)
- How about rank of splits for phylogenetic networks under these models?
- Other ATR models for amino acid substitution?

Open Questions

- EI model 4 states (F81): rank of blocks of flattening?
- EI model, any number of states: change of coordinates (already working on this with G. Dilaver, J. Garbett, R. Homs, A. Korchmaros, N. Paul)

- How about rank of splits for phylogenetic networks under these models?
- Other ATR models for amino acid substitution?

Thanks for your attention!

and thanks to:

- for ATR[4]: Roser Homs, Angelica Torres
- for SAQ, ASAQ: Jesús Fernández-Sánchez, Marina Garrote-López
- for the embedding problem: Jesús Fernández-Sánchez, Jordi Roca-Lacostena

[4] C, Homs, Torres, A novel algebraic approach to time-reversible evolutionary models, SIAM Journal on Applied Mathematics, 2024