

Sufficient statistics for CTMCs: Methods, implementations and applications

Paula Tataru, Asger Hobolth

April 13, 2011

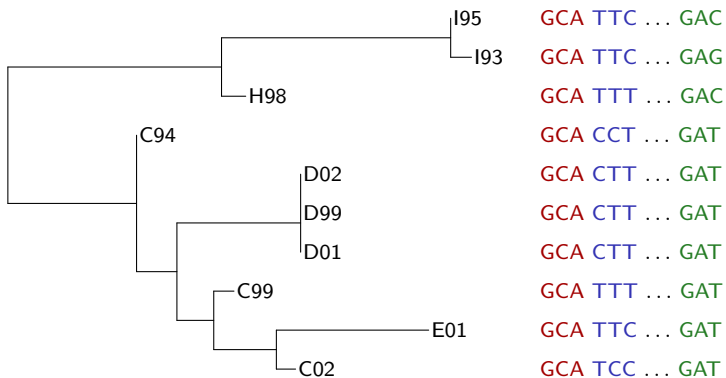
HIV *pol* gene

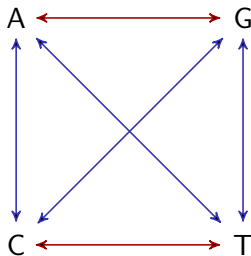
Figure: Phylogeny reconstructed using maximum likelihood

Genetic code

		2 nd base							
		T		C		A		G	
1 st base	T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
		TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys
		TTA	Leu	TCA	Ser	TAA	Stop	TGA	Stop
		TTG	Leu	TCG	Ser	TAG	Stop	TGG	Trp
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	
	CTC	Leu	CCC	Pro	CAC	His	CGC	Arg	
	CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	
	CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	
	ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	
	ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	
	ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg	
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	
	GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly	
	GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly	
	GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly	

Table: Genetic code

Transitions vs transversions



Codon substitution model

Goldman&Yang (94) codon model:

$$q_{ij} = \begin{cases} 0 & \text{if } i \rightarrow j \text{ by more than one substitution} \\ \alpha\kappa\pi_j & \text{if } i \rightarrow j \text{ by a synonymous transition} \\ \alpha\pi_j & \text{if } i \rightarrow j \text{ by a synonymous transversion} \\ \alpha\omega\kappa\pi_j & \text{if } i \rightarrow j \text{ by a non-synonymous transition} \\ \alpha\omega\pi_j & \text{if } i \rightarrow j \text{ by a non-synonymous transversion} \end{cases}$$

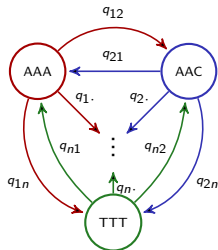
Continuous Time Markov Chains (CTMCs)

- ▶ A CTMC is a stochastic process $\{X(t)|t \geq 0\}$ that fulfils the Markov property.
- ▶ The process is characterized by a rate matrix $Q = (q_{ij})$ that describes the rates at which the process moves from one state to another.

- ▶ Q fulfils $q_{ii} = -\sum_{j \neq i} q_{ij}$.

- ▶ Transition probabilities

$$p_{ij}(t) = \mathbb{P}(X(t) = j | X(0) = i) = (e^{Qt})_{ij}.$$



Likelihood equations

The log likelihood of observing the complete data \mathbf{x} given κ, ω, α is:

$$\begin{aligned} \ell(\kappa, \omega, \alpha; \mathbf{x}) = & -\alpha (\kappa T^{s,ts} + \omega \kappa T^{ns,ts} + \omega T^{ns,tv} + T^{s,tv}) \\ & + N^{ts} \log \kappa + N^{ns} \log \omega + N \log \alpha \end{aligned}$$

Let $\beta = \alpha \kappa$. Maximize likelihood \Leftrightarrow solving a 2nd order equation:

$$\left\{ \begin{array}{l} \frac{\partial l}{\partial \alpha} = -T^{s,tv} - \omega T^{ns,tv} + \frac{N^{tv}}{\alpha} = 0 \\ \frac{\partial l}{\partial \beta} = -T^{s,ts} - \omega T^{ns,ts} + \frac{N^{ts}}{\beta} = 0 \\ \frac{\partial l}{\partial \omega} = -\alpha T^{ns,tv} - \beta T^{ns,ts} + \frac{N^{ns}}{\omega} = 0 \end{array} \right.$$

Missing data problem: EM algorithm

Iterative procedure:

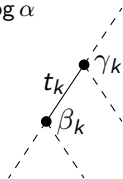
- ▶ expectation step - computes the expectation of the likelihood using the current estimate of the variables;
- ▶ maximization step - estimates the parameters by maximizing the likelihood found in the previous step.

In particular, it can be used to estimate the rate matrix for discretely observed CTMCs.

E step

$$\begin{aligned} \mathbb{E}[\ell(\kappa, \omega, \alpha; \mathbf{x}) | \mathbf{y}] = & -\alpha (\kappa \mathbb{E}[T^{s,ts} | \mathbf{y}] + \omega \kappa \mathbb{E}[T^{ns,ts} | \mathbf{y}] + \omega \mathbb{E}[T^{ns,tv} | \mathbf{y}] + \mathbb{E}[T^{s,tv} | \mathbf{y}]) \\ & + \mathbb{E}[N^{ts} | \mathbf{y}] \log \kappa + \mathbb{E}[N^{ns} | \mathbf{y}] \log \omega + \mathbb{E}[N | \mathbf{y}] \log \alpha \end{aligned}$$

Assuming a fixed phylogeny, we have:



$$\mathbb{E}[T_i | \mathbf{y}] = \sum_{\text{branch } k} \sum_{a,b} \mathbb{P}(\gamma_k = a, \beta_k = b, t_k | \mathbf{y}) \frac{\mathbb{E}[T_i \mathbb{1}(b \text{ after } t_k) | a]}{p_{ab}(t_k)}$$

$$\mathbb{E}[N_{ij} | \mathbf{y}] = \sum_{\text{branch } k} \sum_{a,b} \mathbb{P}(\gamma_k = a, \beta_k = b, t_k | \mathbf{y}) \frac{\mathbb{E}[N_{ij} \mathbb{1}(b \text{ after } t_k) | a]}{p_{ab}(t_k)}$$

Statistics for CTMCs

We are interested in the following summary statistics:

$$\mathbb{E}[N_{cd} \mathbb{1}(X(T) = b) | X(0) = a] = q_{cd} I_{ab}^{cd}(T)$$

where $I_{ab}^{cd}(T) = \int_0^T p_{ac}(t) p_{db}(T-t) dt$.

Sometimes we need sums of expected values:

$$N_{ab}^{\Delta}(T) = \sum_{(c,d) \in \Delta} \mathbb{E}[N_{cd} \mathbb{1}(X(T) = b) | X(0) = a] = \sum_{(c,d) \in \Delta} q_{cd} I_{ab}^{cd}(T)$$

N^{ts} requires $\Delta = \{(c, d) | c \rightarrow d \text{ by a transition}\}$.

Methods

In Hobolth. A & Jensen J.L. (2010), three methods are presented to evaluate the necessary statistics:

1. eigenvalue decomposition - **eigen**;
2. uniformization - **uni**;
3. basic matrix exponentiation - **expm**.

Our aim is to implement them and compare their performance.

1. Eigenvalue decomposition

Let $Q = U\Lambda U^{-1}$

Then
$$\begin{aligned}
 I_{ab}^{cd}(T) &= \int_0^T p_{ac}(t) p_{db}(T-t) dt \\
 &= \int_0^T \left(e^{Qt} \right)_{ac} \left(e^{Q(T-t)} \right)_{db} dt \\
 &= \int_0^T \left(U e^{\Lambda t} U^{-1} \right)_{ac} \left(U e^{\Lambda(T-t)} U^{-1} \right)_{db} dt
 \end{aligned}$$

Setting $J_{ij}(T) = \int_0^T e^{\lambda_i t} e^{\lambda_j(T-t)} dt$

we have $N^\Delta(T) = U \cdot [J(T) \star [U^{-1} \cdot ((\mathbb{1}_\Delta \star Q) \cdot U)]] \cdot U^{-1}$

2. Uniformization

Let $\mu = \max_i (-q_{ii})$ and $R = \frac{1}{\mu}Q + I$

Then
$$\begin{aligned}
 P(t) &= e^{Qt} = e^{(\mu(R-I))t} \\
 &= \sum_{m=0}^{\infty} R^m \frac{(\mu t)^m}{m!} e^{-\mu t} = \sum_{m=0}^{\infty} R^m \text{Pois}(m; \mu t)
 \end{aligned}$$

and
$$N^{\Delta}(T) = \frac{1}{\mu} \sum_{m=0}^{\infty} \text{Pois}(m+1; \mu T) \sum_{l=0}^m R^l \cdot (\mathbb{1}_{\Delta} \star Q) \cdot R^{m-l}$$

Truncation at $s(\lambda)$

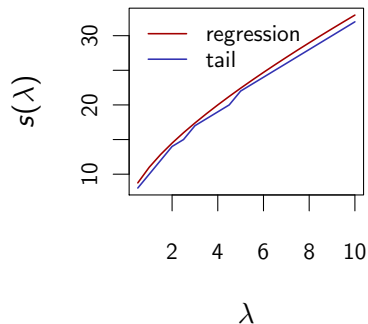
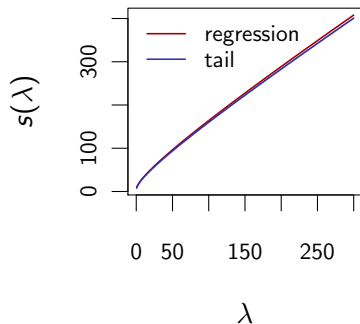


Figure: Comparison between the Poisson tail as determined by **R** and the approximation using linear regression, with $s(\lambda) = 4 + 6\sqrt{\lambda} + \lambda$.

3. Exponentiation

$$\text{Let } A = \begin{bmatrix} Q & \mathbb{1}_\Delta \star Q \\ 0 & Q \end{bmatrix} \Rightarrow e^{At} = \begin{bmatrix} F(T) & G(T) \\ 0 & F(T) \end{bmatrix}$$

$$\text{We have } \frac{d}{dT} e^{AT} = A e^{AT}$$

$$\Rightarrow \begin{bmatrix} F'(T) & G'(T) \\ 0 & F'(T) \end{bmatrix} = \begin{bmatrix} Q & \mathbb{1}_\Delta \star Q \\ 0 & Q \end{bmatrix} \begin{bmatrix} F(T) & G(T) \\ 0 & F(T) \end{bmatrix}$$

$$\Rightarrow G(T) = \int_0^T e^{Qt} (\mathbb{1}_\Delta \star Q) e^{Q(T-t)} dt$$

$$\text{Then } N^\Delta(T) = (e^{AT})_{1:n, (n+1):2n}$$

For exponentiating a matrix, we used the **R** package *expm*.

Expectation time

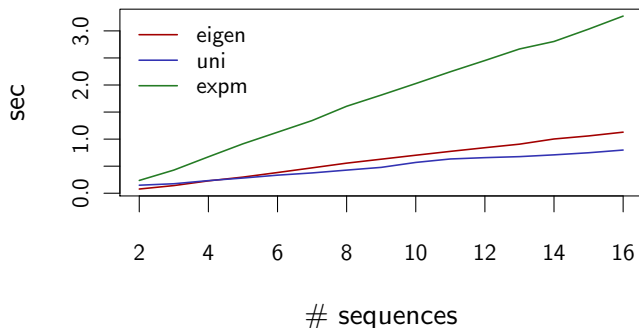


Figure: Time comparison for computing expectations in the EM algorithm

Summary

	$N^\Delta(T)$	accuracy	running time	
			pre	main
eigen	$U \cdot [J(T) \star [U^{-1} \cdot ((\mathbb{1}_\Delta \star Q) \cdot U)]] \cdot U^{-1}$	prone to large errors	$O(n^3)$	$O(n^3)$
uni	$\frac{1}{\mu} \sum_m \text{Pois}(m; \mu T) \sum_{l=0}^{m-1} R^l \cdot (\mathbb{1}_\Delta \star Q) \cdot R^{m-l}$	sum of many small numbers	$O(s(\mu T)n^3)$	$O(s(\mu T)n^3)$
expm	$(e^{AT})_{1:n, (n+1):2n}$	most accurate ¹	none	$O(n^3)$

Table: Summary

¹Higham, J.: *The Scaling and Squaring Method for the Matrix Exponential Revisited* (2003)
Moler, C., Van Loan, C.: *Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later* (2003)