Improving the computational speed of phylogenetic model ranking with data-augmentation-based thermodynamic integration

Nicolas Rodrigue & Stéphane Aris-Brosou
University of Ottawa

Modeling molecular evolution

…as a continuous-time Markov chain

```
A A G A G C T C C T A C T T C G T A A C T...
A T G A G T T G C T A C T T C G G A A C T...
A A G T G T T C C T A C T T C G T A A C T...
A T C A G T T C C A A C T T C G T A A G T...
A T G T G T T C C T T C T T C G T A A C T...
T T G A G T T C C T T C T T C G T C A C T...
T A G A G T T C C T A C T T C G T A A C T...
T T C A G T T C T T A G T A G G T A T C T...
```

…with the 4 nucleotides as state space.

Modeling molecular evolution

…as a continuous-time Markov chain

```
A A G A G C T C C T A C T T C G T A A C T...
A T G A G T T G C T A C T T C G G A A C T...
A A G T G T T C C T A C T T C G T A A C T...
A T C A G T T C C A A C T T C G T A A G T...
A T G T G T T C C T T C T T C G T A A C T...
T T G A G T T C C T T C T T C G T C A C T...
T A G A G T T C C T A C T T C G T A A C T...
T T C A G T T C T T A G T A G G T A T C T...
```

…with the 4 nucleotides as state space.

Modeling molecular evolution

…as a continuous-time Markov chain

$$Q_{ab} \propto \rho_{ab}\pi_b, a \neq b$$

$$Q_{aa} = -\sum_{b \neq a} Q_{ab}$$

$\rho = (\rho_{ab})_{1 \leq a,b \leq 4}$ Relative exchangeability between nucleotides *a* and *b*.

A
C
G
T

$\pi = (\pi_a)_{1 \leq a \leq 4}$ Equilibrium frequency of nucleotide *a*.

A C G T

A C G T

Modeling molecular evolution

…as a continuous-time Markov chain

GTR

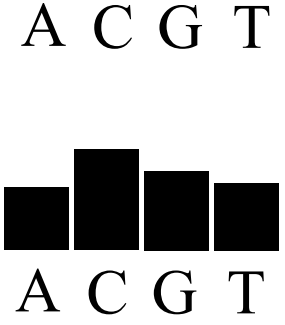
$$Q_{ab} \propto \rho_{ab}\pi_b, a \neq b$$

$$Q_{aa} = -\sum_{b \neq a} Q_{ab}$$

$\rho = (\rho_{ab})_{1 \leq a,b \leq 4}$ Relative exchangeability between nucleotides *a* and *b*.

A
C
G
T

$\pi = (\pi_a)_{1 \leq a \leq 4}$ Equilibrium frequency of nucleotide *a*.



Modeling molecular evolution

…as a continuous-time Markov chain

GTR
HKY
F81
JC

$$Q_{ab} \propto \rho_{ab}\pi_b, a \neq b$$

$$Q_{aa} = -\sum_{b \neq a} Q_{ab}$$

A
C
G
T

$\rho = (\rho_{ab})_{1 \leq a,b \leq 4}$ Relative exchangeability between nucleotides *a* and *b*.

A C G T

$\pi = (\pi_a)_{1 \leq a \leq 4}$ Equilibrium frequency of nucleotide *a*.

A C G T

Modeling molecular evolution

…as a continuous-time Markov chain

Branch length (λ)

G $\qquad\qquad$ A

$$p(A \mid G, \theta) = \int_{\Phi} p(A, \phi \mid G, \theta) d\phi = \left[e^{\lambda Q} \right]_{GA}$$

Given a starting state (G) and the parameters of the Markov process (ρ, φ), the probability of ending up in some other state (A) over a given evolutionary distance (λ, i.e., $\theta = \{\rho, \varphi, \lambda\}$) can be expressed as an integral over all possible substitution mappings (ϕ), and can be computed exploiting a matrix diagonalization routine.

Modeling molecular evolution

…as a continuous-time Markov chain

Branch length (λ)

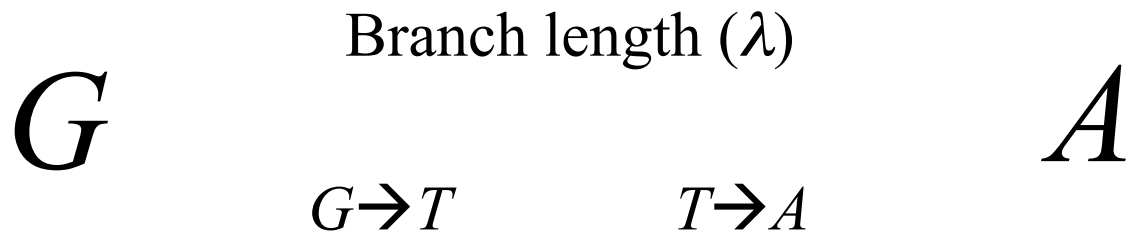
$$G \xrightarrow{G \to A} A$$

$$p(A \mid G, \theta) = \int_{\Phi} p(A, \phi \mid G, \theta) d\phi = \left[e^{\lambda Q} \right]_{GA}$$

Given a starting state (G) and the parameters of the Markov process (ρ, φ), the probability of ending up in some other state (A) over a given evolutionary distance (λ, i.e., $\theta = \{\rho, \varphi, \lambda\}$) can be expressed as an integral over all possible substitution mappings (ϕ), and can be computed exploiting a matrix diagonalization routine.

Modeling molecular evolution

…as a continuous-time Markov chain



$$p(A \mid G,\theta) = \int_{\Phi} p(A,\phi \mid G,\theta)\, d\phi = \left[e^{\lambda Q} \right]_{GA}$$

Given a starting state (G) and the parameters of the Markov process (ρ, φ), the probability of ending up in some other state (A) over a given evolutionary distance (λ, i.e., $\theta = \{\rho, \varphi, \lambda\}$) can be expressed as an integral over all possible substitution mappings (ϕ), and can be computed exploiting a matrix diagonalization routine.

Modeling molecular evolution

…as a continuous-time Markov chain

Branch length (λ)

$$G \qquad \underset{G\rightarrow A \quad A\rightarrow G \qquad G\rightarrow A}{} \qquad A$$

$$p(A \mid G,\theta) = \int_{\Phi} p(A,\phi \mid G,\theta)\, d\phi = \left[e^{\lambda Q} \right]_{GA}$$

Given a starting state (G) and the parameters of the Markov process (ρ, φ), the probability of ending up in some other state (A) over a given evolutionary distance (λ, i.e., $\theta = \{\rho, \varphi, \lambda\}$) can be expressed as an integral over all possible substitution mappings (ϕ), and can be computed exploiting a matrix diagonalization routine.

Modeling molecular evolution

…as a continuous-time Markov chain

Branch length (λ)

G $\qquad\qquad$ A

$$p(A \mid G, \theta) = \int_{\Phi} p(A, \phi \mid G, \theta) d\phi = \left[e^{\lambda Q} \right]_{GA}$$

Given a starting state (G) and the parameters of the Markov process (ρ, φ), the probability of ending up in some other state (A) over a given evolutionary distance (λ, i.e., $\theta = \{\rho, \varphi, \lambda\}$) can be expressed as an integral over all possible substitution mappings (ϕ), and can be computed exploiting a matrix diagonalization routine.

Modeling molecular evolution

…as a continuous-time Markov chain

$$p(D \mid \theta, M) = \prod_i p(D_i \mid \theta, M)$$

```
A A G A G C T C C T A C T T C G T A A C T...
A T G A G T T G C T A C T T C G G A A C T...
A A G T G T T C C T A C T T C G T A A C T...
A T C A G T T C C A A C T T C G T A A G T...
A T G T G T T C C T T C T T C G T A A C T...
T T G A G T T C C T T C T T C G T C A C T...
T A G A G T T C C T A C T T C G T A A C T...
T T C A G T T C T T A G T A G G T A T C T...
```

…with the 4 nucleotides as state space.

Modeling molecular evolution

…as a continuous-time Markov chain

$$p(D \mid \theta, M) = \prod_i p(D_i \mid \theta, M)$$

```
A A G A G C T C C T A C T T C G T A A C T...
A T G A G T T G C T A C T T C G G A A C T...
A A G T G T T C C T A C T T C G T A A C T...
A T C A G T T C C A A C T T C G T A A G T...
A T G T G T T C C T T C T T C G T A A C T...
T T G A G T T C C T T C T T C G T C A C T...
T A G A G T T C C T A C T T C G T A A C T...
T T C A G T T C T T A G T A G G T A T C T...
```

…with the 61 sense codons as its state space.

Modeling molecular evolution

Yang & Nielsen Mol. Biol. Evol., 25:568-579, 2008.

The instantaneous rate of substitution from codon *m* to codon *n* is proportional to:

$$Q_{mn} \propto \begin{cases} \rho_{m_j n_j} \pi_{n_j} & \text{if } m \text{ and } n \text{ are synonymous, and differ only at codon position } j, \\ \rho_{m_j n_j} \pi_{n_j} \dfrac{\ln\left(\psi_{f(n)} / \psi_{f(m)}\right)}{1 - e^{-\ln\left(\psi_{f(n)} / \psi_{f(m)}\right)}} & \text{if } m \text{ and } n \text{ are nonsynonymous, and differ only at codon position } j, \\ 0 & \text{otherwise.} \end{cases}$$

Here, $f(n)$ returns an index (from 1 to 20) of the amino acid encoded by codon *n*, and all sites of the alignment have the same amino acid profile (ψ).

Modeling molecular evolution

Halpern & Bruno Mol. Biol. Evol., 15:910-917, 1998.

For site *i*, the instantaneous rate of substitution from codon *m* to codon *n* is proportional to:

$$Q_{mn}^{(i)} \propto \begin{cases} \rho_{m_j n_j} \pi_{n_j} & \text{if } m \text{ and } n \text{ are synonymous, and differ only at codon position } j, \\ \rho_{m_j n_j} \pi_{n_j} \dfrac{\ln\left(\psi_{f(n)}^{(i)} / \psi_{f(m)}^{(i)}\right)}{1 - e^{-\ln\left(\psi_{f(n)}^{(i)} / \psi_{f(m)}^{(i)}\right)}} & \text{if } m \text{ and } n \text{ are nonsynonymous, and differ only at codon position } j, \\ 0 & \text{otherwise.} \end{cases}$$

In this case, all sites of the alignment have their own amino acid profile.

Capturing heterogeneous amino acid selective pressures

The model proposed by Yang & Nielsen (MBE, 25:568-579, 2008) is homogeneous with respect to amino acid frequencies (analogous to having a single ψ vector).

Capturing heterogeneous amino acid selective pressures

The model proposed by Yang & Nielsen (MBE, 25:568-579, 2008) is homogeneous with respect to amino acid frequencies (analogous to having a single ψ vector).

The model proposed by Halpern & Bruno (MBE, 15:910-917, 1998) would have each codon site with its own amino acid context (analogous to having as many ψ vectors as there are codon sites in the alignment).

Capturing heterogeneous amino acid selective pressures

The model proposed by Yang & Nielsen (MBE, 25:568-579, 2008) is homogeneous with respect to amino acid frequencies (analogous to having a single ψ vector).

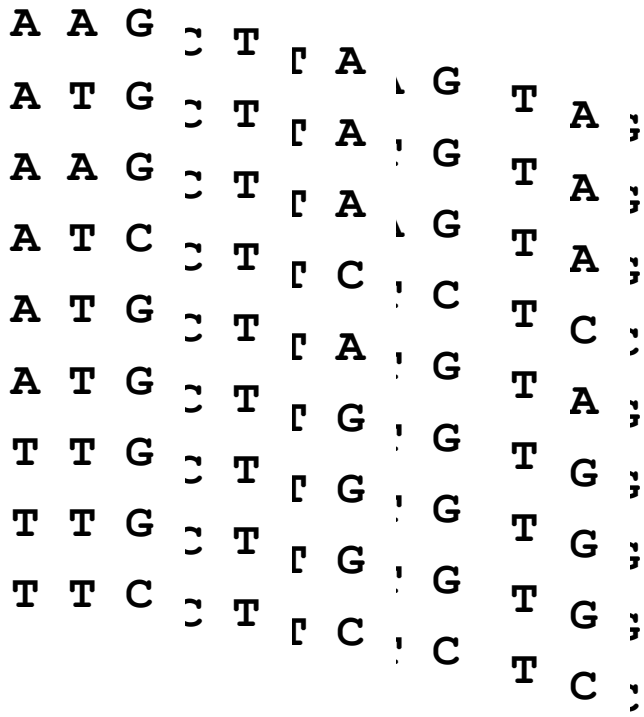
The model proposed by Halpern & Bruno (MBE, 15:910-917, 1998) would have each codon site with its own amino acid context (analogous to having as many ψ vectors as there are codon sites in the alignment).

We have recently proposed mixture modeling approaches (Rodrigue et al., PNAS, 107:4629-4634, 2010) which provide a compromise between YN and HB.

Capturing heterogeneous amino acid selective pressures

Instead of viewing all data columns as arising from a single global model, as in standard homogeneous models…

Generative model



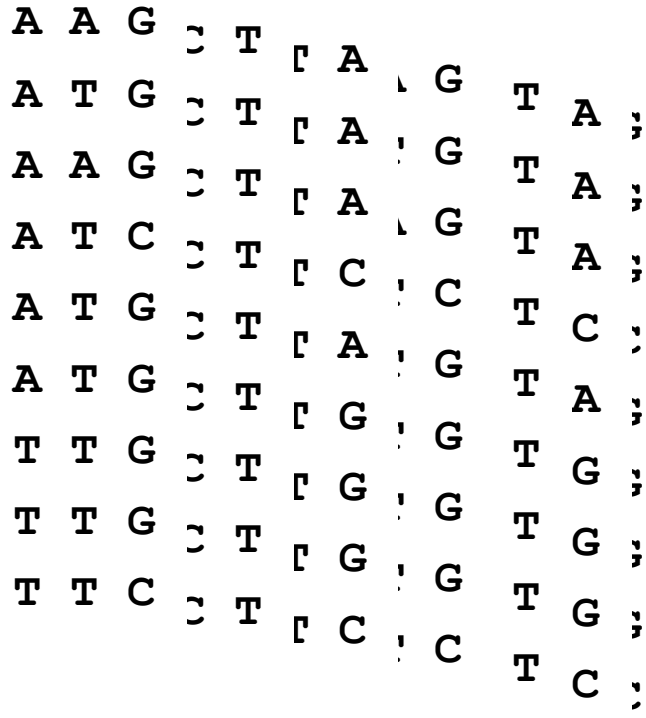
Capturing heterogeneous amino acid selective pressures

Instead of viewing all columns of the data as arising from a single global model, as in standard homogeneous models, in a mixture modeling approach, each column is viewed a arising from one of several possible models.

Generative model

Generative model

Generative model



Capturing heterogeneous amino acid selective pressures

Instead of viewing all columns of the data as arising from a single global model, as in standard homogeneous models, in a mixture modeling approach, each column is viewed a arising from one of several possible models.

For computation simplicity, we use an empirical mixture model based on either 20, 40, or 60 components, inferred from large data sets (Li, Gascuel, & Lartillot, Bioinformatics, 2008). The only additional degrees of freedom in these mixture models are the weights of the different components, which we treat as free parameters.

The Bayesian paradigm

likelihood function prior

$$p(\theta \mid D, M) = \frac{p(D \mid \theta, M)\, p(\theta \mid M)}{p(D \mid M)}$$

posterior

marginal likelihood

D Data set

θ Parameter vector

M Model

The Bayesian paradigm

likelihood function prior

$$p(\theta \mid D, M) = \frac{p(D \mid \theta, M)\, p(\theta \mid M)}{p(D \mid M)}$$

posterior

marginal likelihood

Obtain a sample from the posterior distribution using Metropolis-Hastings and Gibbs sampling algorithms (Markov chain Monte Carlo).

The Bayesian paradigm

When comparing two models M_0 and M_1, one can look at the ratio of their respective marginal likelihoods. This ratio is known as the Bayes factor B.

$$B = \frac{p(D \mid M_1)}{p(D \mid M_0)}$$

$$p(D \mid M) = \int_{\Theta} p(D \mid \theta, M) p(\theta \mid M) d\theta$$

The Bayesian paradigm

When comparing two models M_0 and M_1, one can look at the ratio of their respective marginal likelihoods. This ratio is known as the Bayes factor B.

$$B = \frac{p(D \mid M_1)}{p(D \mid M_0)}$$

$$p(D \mid M) = \int_{\Theta} p(D \mid \theta, M)\, p(\theta \mid M)\, d\theta$$

likelihood function

The Bayesian paradigm

When comparing two models M_0 and M_1, one can look at the ratio of their respective marginal likelihoods. This ratio is known as the Bayes factor B.

$$B = \frac{p(D \mid M_1)}{p(D \mid M_0)}$$

$$p(D \mid M) = \int_{\Theta} p(D \mid \theta, M)\, p(\theta \mid M)\, d\theta$$

likelihood function prior

The Bayesian paradigm

When comparing two models M_0 and M_1, one can look at the ratio of their respective marginal likelihoods. This ratio is known as the Bayes factor B.

$$B = \frac{p(D \mid M_1)}{p(D \mid M_0)}$$

$B > 1$ is considered as *evidence* for M_1.

The Bayesian paradigm

When comparing two models M_0 and M_1, one can look at the ratio of their respective marginal likelihoods. This ratio is known as the Bayes factor B.

$$B = \frac{p(D \mid M_1)}{p(D \mid M_0)}$$

Bayes factors can be reliably computed in phylogenetic contexts using model-switch thermodynamic integration (Lartillot & Philippe, Syst. Biol., 2006).

The Bayesian paradigm

Problem: The model-switch method is CPU intensive.

Evaluating the mutation-selection codon substitution model of Yang & Nielsen (2008) requires weeks on a modern desktop computer, and evaluating the site-specific model proposed by Halpern & Bruno (1998) in intractable.

The Bayesian paradigm

Problem: The model-switch method is CPU intensive.

Evaluating the mutation-selection codon substitution model of Yang & Nielsen (2008) requires weeks on a modern desktop computer, and evaluating the site-specific model proposed by Halpern & Bruno (1998) in intractable.

Solution: Make improvements to the model-switch.

Fan et al. (Syst. Biol., in press, 2011) and Xie et al. (MBE, in press, 2011) have recently proposed improvements. But these efforts don't take advantage of a recent computation development: **data-augmentation**.

Phylogenetic data-augmentation

Decomposing the likelihood function as an integral over all substitution mappings (data augmentation/auxiliary variables)

$$p(D \mid \theta, M) = \int_{\Phi} p(D, \phi \mid \theta, M) d\phi$$

```
A A G A G C T C C T A C T T C G T A A C T...
A T G A G T T G C T A C T T C G G A A C T...
A A G T G T T C C T A C T T C G T A A C T...
A T C A G T T C C A A C T T C G T A A G T...
A T G T G T T C C T T C T T C G T A A C T...
T T G A G T T C C T T C T T C G T C A C T...
T A G A G T T C C T A C T T C G T A A C T...
T T C A G T T C T T A G T A G G T A T C T...
```

Phylogenetic data-augmentation

Decomposing the likelihood function as an integral over all substitution mappings (data augmentation/auxiliary variables)

$$p(D \mid \theta, M) = \int_{\Phi} p(D, \phi \mid \theta, M) d\phi$$

```
A A G A G C T C C T A C T T C G T A A C T...
A T G A G T T G C T A C T T C G G A A C T...
A A G T G T T C C T A C T T C G T A A C T...
A T C A G T T C C A A C T T C G T A A G T...
A T G T G T T C C T T C T T C G T A A C T...
T T G A G T T C C T T C T T C G T C A C T...
T A G A G T T C C T A C T T C G T A A C T...
T T C A G T T C T T A G T A G G T A T C T...
```

Nielsen, R. Syst. Biol., 51:729-739, 2002.

Rodrigue *et al*. Bioinformatics, 24:56-62, 2008.

Phylogenetic data-augmentation

Decomposing the likelihood function as an integral over all substitution mappings (data augmentation/auxiliary variables)

$$p(D \mid \theta, M) = \int_{\Phi} p(D, \phi \mid \theta, M) d\phi$$

```
A A G A G C T C C T A C T T C G T A A C T...
A T G A G T T G C T A C T T C G G A A C T...
A A G T G T T C C T A C T T C G T A A C T...
A T C A G T T C C A A C T T C G T A A G T...
A T G T G T T C C T T C T T C G T A A C T...
T T G A G T T C C T T C T T C G T C A C T...
T A G A G T T C C T A C T T C G T A A C T...
T T C A G T T C T T A G T A G G T A T C T...
```

Nielsen, R. Syst. Biol., 51:729-739, 2002.

Rodrigue *et al*. Bioinformatics, 24:56-62, 2008.

Phylogenetic data-augmentation

Decomposing the likelihood function as an integral over all substitution mappings (data augmentation/auxiliary variables)

$$p(D \mid \theta, M) = \int_{\Phi} p(D, \phi \mid \theta, M) d\phi$$

```
A A G A G C T C C T A C T T C G T A A C T...
A T G A G T T G C T A C T T C G G A A C T...
A A G T G T T C C T A C T T C G T A A C T...
A T C A G T T C C A A C T T C G T A A G T...
A T G T G T T C C T T C T T C G T A A C T...
T T G A G T T C C T T C T T C G T C A C T...
T A G A G T T C C T A C T T C G T A A C T...
T T C A G T T C T T A G T A G G T A T C T...
```

Nielsen, R. Syst. Biol., 51:729-739, 2002.

Rodrigue *et al*. Bioinformatics, 24:56-62, 2008.

Phylogenetic data-augmentation

Decomposing the likelihood function as an integral over all substitution mappings (data augmentation/auxiliary variables)

$$p(D \mid \theta, M) = \int_{\Phi} p(D, \phi \mid \theta, M) d\phi$$

```
A A G A G C T C C T A C T T C G T A A C T...
A T G A G T T G C T A C T T C G G A A C T...
A A G T G T T C C T A C T T C G T A A C T...
A T C A G T T C C A A C T T C G T A A G T...
A T G T G T T C C T T C T T C G T A A C T...
T T G A G T T C C T T C T T C G T C A C T...
T A G A G T T C C T A C T T C G T A A C T...
T T C A G T T C T T A G T A G G T A T C T...
```

Nielsen, R. Syst. Biol., 51:729-739, 2002.

Rodrigue *et al*. Bioinformatics, 24:56-62, 2008.

Phylogenetic data-augmentation

likelihood function prior

$$p(\theta \mid D, M) = \frac{p(D \mid \theta, M)\, p(\theta \mid M)}{p(D \mid M)}$$

posterior

marginal likelihood

D Data set

θ Parameter vector

M Model

Phylogenetic data-augmentation

augmented likelihood prior

$$p(\theta,\phi \mid D,M) = \frac{p(D,\phi \mid \theta,M)\,p(\theta \mid M)}{p(D \mid M)}$$

joint posterior

marginal likelihood

Phylogenetic data-augmentation

augmented likelihood prior

$$p(\theta,\phi \mid D,M) = \frac{p(D,\phi \mid \theta,M)\,p(\theta \mid M)}{p(D \mid M)}$$

joint posterior

marginal likelihood

The θ component of the above joint distribution is distributed as it would be under the canonical posterior. But, in practice, we can make a sampler based on the joint distribution that is much faster than a sampler that explicitly integrates out the auxiliary variable ϕ.

Computational improvement

Natural log Bayes factors using MG codon substitution model as a reference, computed on the Globin17-144 data set.

Model	Classical thermo	DA-thermo
MG-MutSelYN	44.2 (1) [~2 weeks]	44.3 (0.31) [~1 hour]
MG-MutSelC20	236.3 (1) [~3 months]	236.3 (0.35) [~5 hours]
MG-MutSelC40	256.9 (1) [~3.5 months]	256.8 (0.38) [~7 hours]
MG-MutSelC60	269.1 (1) [~4 months]	269.0 (0.41) [~9 hours]
MG-MutSelHB	NA	159.3 (0.51) [~12 hours]

Computational improvement

Natural log Bayes factors using MG codon substitution model as a reference, computed on the Globin17-144 data set.

Model	Classical thermo	DA-thermo
MG-MutSelYN	**44.2** (1) [~2 weeks]	**44.3** (0.31) [~1 hour]
MG-MutSelC20	**236.3** (1) [~3 months]	**236.3** (0.35) [~5 hours]
MG-MutSelC40	**256.9** (1) [~3.5 months]	**256.8** (0.38) [~7 hours]
MG-MutSelC60	**269.1** (1) [~4 months]	**269.0** (0.41) [~9 hours]
MG-MutSelHB	NA	159.3 (0.51) [~12 hours]

Computational improvement

Natural log Bayes factors using MG codon substitution model as a reference, computed on the Globin17-144 data set.

Model	Classical thermo	DA-thermo
MG-MutSelYN	44.2 (**1**) [~2 weeks]	44.3 (**0.31**) [~1 hour]
MG-MutSelC20	236.3 (**1**) [~3 months]	236.3 (**0.35**) [~5 hours]
MG-MutSelC40	256.9 (**1**) [~3.5 months]	256.8 (**0.38**) [~7 hours]
MG-MutSelC60	269.1 (**1**) [~4 months]	269.0 (**0.41**) [~9 hours]
MG-MutSelHB	NA	159.3 (0.51) [~12 hours]

Computational improvement

Natural log Bayes factors using MG codon substitution model as a reference, computed on the Globin17-144 data set.

Model	Classical thermo	DA-thermo
MG-MutSelYN	44.2 (1) **[~2 weeks]**	44.3 (0.31) **[~1 hour]**
MG-MutSelC20	236.3 (1) **[~3 months]**	236.3 (0.35) **[~5 hours]**
MG-MutSelC40	256.9 (1) **[~3.5 months]**	256.8 (0.38) **[~7 hours]**
MG-MutSelC60	269.1 (1) **[~4 months]**	269.0 (0.41) **[~9 hours]**
MG-MutSelHB	**NA**	159.3 (0.51) **[~12 hours]**

Computational improvement

Natural log Bayes factors using MG codon substitution model as a reference, computed on the Globin17-144 data set.

Model	Classical thermo	DA-thermo
MG-MutSelYN	44.2 (1) [~2 weeks]	44.3 (0.31) [~1 hour]
MG-MutSelC20	236.3 (1) [~3 months]	236.3 (0.35) [~5 hours]
MG-MutSelC40	256.9 (1) [~3.5 months]	256.8 (0.38) [~7 hours]
MG-MutSelC60	269.1 (1) [~4 months]	269.0 (0.41) [~9 hours]
MG-MutSelHB	NA	159.3 (0.51) [~12 hours]

Concluding message

In some Monte Carlo contexts, introducing variables that are not really needed (i.e., that we know how to integrate out of the probability calculations) can render approximations both much faster and more precise.

With the many probabilistic modeling possibilities arising in biodiversity science, we will need to evaluate big integrals (multidimensional, non-analytical). We should keep in mind that there are always other methods of approximation, some of which can be much faster than others.

Acknowledgments

uOttawa

L'Université canadienne
Canada's university

NSERC
CRSNG

CENTRE DE LA SCIENCE DE LA BIODIVERSITÉ DU QUÉBEC
QUEBEC CENTRE FOR BIODIVERSITY SCIENCE