Using Non-Reversible Context-Dependent Evolutionary Models to Study Substitution Patterns in Primate Non-Coding Sequences

Guy Baele · Yves Van de Peer · Stijn Vansteelandt

Received: 8 December 2009 / Accepted: 26 May 2010
© Springer Science+Business Media, LLC 2010

Abstract We discuss the importance of non-reversible evolutionary models when analyzing context-dependence. Given the inherent non-reversible nature of the well-known CpG-methylation-deamination process in mammalian evolution, non-reversible context-dependent evolutionary models may be well able to accurately model such a process. In particular, the lack of constraints on non-reversible substitution models might allow for more accurate estimation of context-dependent substitution parameters. To demonstrate this, we have developed different time-homogeneous context-dependent evolutionary models to analyze a large genomic dataset of primate ancestral repeats based on existing independent evolutionary models. We have calculated the difference in model fit for each of these models using Bayes Factors obtained via thermodynamic integration. We find that non-reversible context-dependent models can drastically increase model fit when compared to independent models and this on two primate non-coding datasets. Further, we show that further improvements are possible by clustering similar parameters across contexts.

Keywords Context-dependent evolution · Nearest-neighbor influences · Context effect · Bayes Factor · Thermodynamic integration

Abbreviations

GTR General time-reversible
BF Bayes Factor
QE Quasistatic estimate

Introduction

In many studies on molecular evolution, the independent general time-reversible model (GTR; Lanavé et al. 1984) is used as the model of choice due to its ability to incorporate a range of different substitution behaviors (e.g., Yang 1994). However, context-dependent evolutionary models and the accompanying methodologies to handle the increased model complexity have been the subject of an increasing number of studies (e.g., Baele et al. 2008, 2010, for a review). These studies demonstrate that the assumption of site-independent evolution is overly restrictive and that evidence for context-dependent substitution patterns is abundant in large mammalian datasets. Context-dependent models are often parameter-rich which enables many different model compositions.

In this article, we develop and test nine different non-reversible context-dependent models as extensions of popular independent models and compare them to the performance of two previously introduced context-dependent models (Baele et al. 2008, 2010) and the context-dependent model of Hwang and Green (2004). Specifically, we extend the following non-reversible models: the general

G. Baele · Y. Van de Peer (✉)
Department of Plant Systems Biology, VIB, Ghent University,
Technologiepark 927, 9052 Ghent, Belgium
e-mail: yves.vandepeer@psb.vib-ugent.be

G. Baele
e-mail: gubae@psb.vib-ugent.be

G. Baele · Y. Van de Peer
Bioinformatics and Evolutionary Genomics, Department of
Molecular Genetics, Ghent University, 9052 Ghent, Belgium

S. Vansteelandt
Department of Applied Mathematics and Computer Science,
Ghent University, Krijgslaan 281 S9, 9000 Ghent, Belgium
e-mail: stijn.vansteelandt@ugent.be

non-reversible (GMR) model (e.g., Huelsenbeck et al. 2002), the evolutionary model of Schadt et al. (1998) assuming two alternate prior distributions which may influence estimation of the model's parameters (e.g., Zwickl and Holder 2004), the model of Takahata and Kimura (1981), the two models described in Gojobori et al. (1982), the model of Tajima and Nei (1984) and the model of Blaisdell (1985). Finally, we also present a symmetrical non-reversible context-dependent model of evolution. These models are assumed to be time-homogeneous, i.e., we assume that the rate matrices do not change over time.

Our previous study has focused on developing an efficient methodology to analyze datasets using context-dependent models and on performing model selection using thermodynamic integration. Specifically, we have introduced a framework to study context-dependent evolutionary patterns and assess the difference in model fit, through thermodynamic integration, brought about by a given context-dependent model, which assumed independence at the ancestral root sequence (Baele et al. 2008). Given the increase in the number of parameters indexing such context-dependent models, we have proposed two clustering strategies for parameter reduction, along with an approach to split the calculation of Bayes Factors using thermodynamic integration over multiple processors so as to reduce computation time (Baele et al. 2009). Recently, we have studied the importance of various ancestral root distributions accompanying context-dependent models and their influence on accurately estimating context-dependent substitution patterns (Baele et al. 2010).

In this article, we first analyze a large dataset of Ancestral Repeat sequences which we have extensively described in a previous study (Baele et al. 2008) and go on to analyze a pseudogenes dataset that has been analyzed previously by among others Miyamoto et al. (1987) and Yang (1994), albeit only using independent evolutionary models. Pseudogenes are commonly encountered in the genomes of many organisms and are characterized by close similarities to one or more paralogous genes, yet they are non-functional (Mighell et al. 2000). Pseudogenes are thought to arise by tandem duplication of genes, with ensuing loss of function as a result of gradual accumulation of disabling mutations (for a review, see Balakirev and Ayala 2003). Even though pseudogenes are often described by the derogatory term “junk DNA,” they can provide a wealth of information concerning evolutionary ancestry (e.g., Yang 1994), as we will describe in this article. The independent HKY (Hasegawa et al. 1985) and GTR (Lanavé et al. 1984) models were used in Yang's work (1994) to analyze these pseudogene sequences, with both being concluded to be good candidates for describing the evolution of those pseudogene sequences. The authors also found that assuming gamma-distributed rates across sites

did not lead to an extremely significant improvement over a model assuming a single rate.

The analyses of Yang (1994) have greatly contributed to the current knowledge of evolutionary models and model testing. In this article, we perform thorough analyses of two primate datasets and generate results that do not concur with the conclusion reported in the work of Yang (1994), i.e., that the use of the GTR model is recommended especially for large datasets or for sequences with extreme substitution patterns and that the use of an unrestricted model is not worthwhile. Specifically, we demonstrate that small datasets (datasets that may not have sufficient sites to accurately estimate model parameters, especially in the case of context-dependent models where sufficient sites with specific neighbours are needed to estimate each context-dependent set of model parameters) do not necessarily call for simple (independent) evolutionary models but that such datasets may carry strong evidence for complex context-dependent substitution patterns.

We conclude that the two non-coding primate datasets analyzed in this article exhibit strong context-dependent effects, with a clear presence of the CpG-methylation-deamination process (e.g., Fryxell and Zuckerkandl 2000) established in both. Even though we demonstrate that context-dependent models drastically increase model fit for these two datasets, we also present a clustering strategy to show that these parameter-rich models can be further optimized (thereby increasing model fit even more). This strategy focuses on clustering parameters that are similar across multiple contexts.

Materials and Methods

Data

We have used a first dataset consisting of 10 vertebrate species (Human, *Homo Sapiens*; Chimpanzee, *Pan Troglodytes*; Gorilla, *Gorilla Gorilla*; Orang-utan, *Pongo Pygmaeus*; Baboon, *Papio Anubis*; Macaque, *Macaca Mulatta*; Vervet, *Cercopithecus Aethiops*; Marmoset, *Callithrix jacchus*; Dusky Titi, *Callicebus Moloch*; and Squirrel Monkey, *Saimiri Sciureus*). This dataset is a subset of the alignment analyzed in the work of Margulies et al. (2006), containing sequences all orthologous to a ~1.9-Mb region on human chromosome 7q31.3 which contains the cystic fibrosis transmembrane conductance regulator gene (CFTR). The original dataset was aligned using TBA (Blanchette et al. 2004), of which we have taken a subset using *maf_order* (Human as reference sequence, this avoids rerunning TBA with this subset of 10 sequences; see <http://bio.cse.psu.edu/>). Ancestral repeats in the human sequence were detected using RepeatMasker (Smit et al.

1996–2004) with the RepBase Update libraries (September 2007; Jurka 2000). Simple repeats, low complexity regions, members of the Alu family, RNA elements that diverged less than 25% and L1 elements that diverged less than 20% from the reconstructed ancestral sequence were removed using an adaptation of the script by Elliot Margulies (Margulies et al. 2003). The resulting coordinates were mapped onto our alignment, after which the ancestral repeat sequences were extracted, resulting in a dataset consisting of 114,726 sites for each of the 10 sequences. We use the same fixed consensus tree as in previous work (Baele et al. 2008) and refer to this dataset as the “Ancestral Repeats” dataset.

A second dataset consists of the $\psi\eta$ -globin pseudogene sequences of six primates (Human, *Homo Sapiens*; Chimpanzee, *Pan Troglodytes*; Gorilla, *Gorilla Gorilla*; Orang-utan, *Pongo Pygmaeus*; Rhesus Monkey, *Macaca Mulatta*; and Spider Monkey, *Ateles Geoffroyi*), containing 6,166 nucleotides in each sequence. We have used the fixed consensus tree shown in the work of Yang (1994) and we refer to this dataset as the “Pseudogenes” dataset.

Evolutionary Models

We have tested a number of context-dependent evolutionary models that are derived from frequently used independent models. As there are 16 neighboring base combinations (also known as contexts), we used 16 identical model structures for our context-dependent analyses. Given that this article is aimed at gaining a thorough understanding of the evolutionary patterns in mammalian, and more specifically primate, sequences we have focused our attention on non-reversible evolutionary models due to the presence of the CpG-methylation-deamination process in such sequences. Row and column character states for each of the model are in the following order: A, G, C, and T. For the description of each model, we have used commonly used notations found in the literature. In discussing the inferred parameter values (i.e., the instantaneous rates of change), we will use the rXY notation, with X the starting base and Y the ending base over a branch. We will also use the following notation to clarify the influence of the neighboring bases in the case of context-dependent models: AXG represents a site X under consideration, flanking with a base A as the 5' neighbor and a base G as the 3' neighbor. We also use the terms “independent,” “site-independent,” and “context-independent” as synonyms throughout the text.

The GNR model is the most parameter-rich model we have used, consisting of 12 instantaneous rates of change in each model:

$$Q = \{q_{ij}\} = \begin{pmatrix} - & a & b & c \\ g & - & d & e \\ h & i & - & f \\ j & k & l & - \end{pmatrix}.$$

The diagonals of this rate matrix are specified by the requirement that the rows sum to 0 (e.g., Huelsenbeck et al. 2002). As mentioned before, we do not allow this instantaneous rate matrix to change over time, i.e., the evolutionary process is time-homogeneous. As the Q matrix is non-reversible, the set(s) of (context-dependent) base frequencies π must be calculated by solving the equations $\pi Q = 0$. Using this GNR model in a context-dependent framework results in 192 instantaneous rates of change (12 for each of the 16 neighboring base combinations) for the evolutionary model alone (i.e., not including the parameters describing the ancestral root distribution) and 64 base (or model) frequencies (16 sets of 4 context-dependent frequencies).

In the interest of balancing a parameter-performance trade-off, we have developed a non-reversible evolutionary model which is aimed at capturing complementary substitutions within a given context. This model uses half the number of parameters when compared to the GNR model, both in its independent and in context-dependent version and has the following structure:

$$Q = \{q_{ij}\} = \begin{pmatrix} - & a & b & c \\ f & - & d & e \\ e & d & - & f \\ c & b & a & - \end{pmatrix}.$$

Schadt et al. (1998) introduced a non-reversible model which allows for easy analytical calculation of the base frequencies and contains eight different instantaneous rates of change:

$$Q = \{q_{ij}\} = \begin{pmatrix} - & \alpha & \gamma & \lambda \\ \varepsilon & - & \gamma & \lambda \\ \delta & \kappa & - & \beta \\ \delta & \kappa & \sigma & - \end{pmatrix}.$$

Using Schadt et al.'s non-reversible model in a context-dependent framework results in 128 different instantaneous rates of change for the evolutionary model. Compared to the GNR model, this model assumes that rAC equals rGC, rAT equals rGT, rCA equals rTA, and rCG equals rTG. This assumption is also made for each neighboring base combination in our context-dependent analyses.

Takahata and Kimura (1981) use only five different instantaneous rates of change to describe evolutionary patterns. This model drastically decreases the total number of parameters in our context-dependent analyses to 80, and is additionally convenient for its analytical properties for fast calculation:

$$Q = \{q_{ij}\} = \begin{pmatrix} - & \alpha & \delta & \gamma \\ \beta & - & \gamma & \varepsilon \\ \varepsilon & \gamma & - & \beta \\ \gamma & \delta & \alpha & - \end{pmatrix}.$$

Vis-à-vis the GNR model, this model assumes that $r_{AT} = r_{GC} = r_{CG} = r_{TA}$, $r_{AG} = r_{TC}$, $r_{GA} = r_{CT}$, $r_{AC} = r_{TG}$, and $r_{CA} = r_{GT}$. While such properties may be reasonable when analyzing sequences in an independent manner, they are possibly restrictive in context-dependent analyses.

Gojobori et al. (1982) introduced a 4-parameter and a 6-parameter model of nucleotide evolution, which leads to context-dependent models with respectively 64 and 96 different instantaneous rates of change. The 4-parameter model has the following structure:

$$Q = \{q_{ij}\} = \begin{pmatrix} - & \alpha & \theta\alpha & \gamma \\ \beta & - & \gamma & \theta\beta \\ \theta\beta & \gamma & - & \beta \\ \gamma & \theta\alpha & \alpha & - \end{pmatrix}$$

and the 6-parameter model looks as follows:

$$Q = \{q_{ij}\} = \begin{pmatrix} - & \alpha & \alpha & \gamma \\ \beta & - & \eta & \beta \\ \beta & \delta & - & \beta \\ \varepsilon & \alpha & \alpha & - \end{pmatrix}.$$

Tajima and Nei (1984) designed their model to resolve the biased estimates in methods for estimating the number of nucleotide substitutions developed by among other Takahata and Kimura (1981). Their model contains 4 different instantaneous rates of change to describe evolution:

$$Q = \{q_{ij}\} = \begin{pmatrix} - & b & c & d \\ a & - & c & d \\ a & b & - & d \\ a & b & c & - \end{pmatrix}.$$

Compared to the GNR model, this model assumes that $r_{GA} = r_{CA} = r_{TA}$, $r_{AG} = r_{CG} = r_{TG}$, $r_{AC} = r_{GC} = r_{TC}$ and $r_{AT} = r_{GT} = r_{CT}$. This results in 64 parameter for our context-dependent version of the model of Tamura and Nei (1984).

Blaisdell (1985) designed his evolutionary model following the approaches of Jukes and Cantor (1969), Holmquist (1976; not discussed here), Kimura (1980, 1981; not discussed here), Takahata and Kimura (1981), Gojobori et al. (1982), Lanavé et al. (1984; not discussed here), and Tamura and Nei (1984). The model has four different instantaneous rates of change (hence 64 in its context-dependent version):

$$Q = \{q_{ij}\} = \begin{pmatrix} - & a & d & d \\ c & - & d & d \\ b & b & - & c \\ b & b & a & - \end{pmatrix}.$$

In previous work (Baele et al. 2008), we have introduced a context-dependent model based on the GTR model

(Lanavé et al. 1984), which we denote as GTR16C. The independent evolutionary model contains six rate exchangeability parameters and four base frequencies, leading to 96 rate exchangeability parameters for its context-dependent counterpart. This amounts to 192 different instantaneous rates of change in the context-dependent version of the model. The GTR model has the following structure:

$$\begin{pmatrix} - & \pi_G\alpha & \pi_C\beta & \pi_T\gamma \\ \pi_A\alpha & - & \pi_C\delta & \pi_T\varepsilon \\ \pi_A\beta & \pi_G\delta & - & \pi_T\eta \\ \pi_A\gamma & \pi_G\varepsilon & \pi_C\eta & - \end{pmatrix}.$$

At the time, we only used a single set of base frequencies to describe the sequence composition. In other words, the same base frequencies were used in each of the 16 models corresponding to the 16 neighboring base combinations (i.e., only four base frequencies for the complete context-dependent model). As the parameter values differed according to the context, this led to different substitution patterns depending on the context. In a follow-up study (Baele et al. 2010), we relaxed this restrictive assumption to introduce the concept of context-dependent base (or equilibrium) frequencies, leading to 16 sets of four base frequencies, one for each neighboring base combination. Even though this model is reversible, its context-dependent analog is not due to the fact that the base frequencies are context-dependent as well.

The context-dependent models mentioned above essentially partition the data such that different models of evolution are assigned to different sites, based on the two immediate neighbours of those sites. As opposed to such a data partitioning approach, different models can also be applied with varying probabilities to the same site in an alignment. Each of these models may then be weighted by its probability of occurrence, which can be estimated from the data (Pagel and Meade 2004). Such an approach is site-independent and fits two or more qualitatively different models of sequence evolution to each site in an alignment, without specifying in advance the nature of the models, their relative probabilities, and without having knowledge of which sites are best fit by which model. The model of Pagel and Meade (PM04; 2004) uses multiple instances of the GTR model with a common set of base frequency parameters across all rate matrices.

As argued by Steel (2005), evolutionary models should reflect biological realism and hence be designed with the sequence data in mind. Hwang and Green (2004) developed a context-dependent model, which in its most parameter-rich form closely resembles that of 16 GNR models. However, the authors proposed to cut the number of parameters in half due to complementary substitutions present in the greater cystic fibrosis transmembrane

regulator region (CFTR) they were studying. This allowed the model to consist of 96 instead of 192 different instantaneous rates of change to describe mammalian evolution. The authors also allowed for lineage-dependent substitution patterns, which is out of the scope of this research article.

Bayesian MCMC Methodology

In previous studies (Baele et al. 2008, 2009, 2010), we have established a Bayesian Markov chain Monte Carlo (MCMC) methodology that allows for estimating context-dependent substitution patterns on a fixed tree topology, as well as estimating the ancestral root distribution parameters for various orders of Markov chains. This methodology comprises calculation of the full data likelihood through the use of a data augmentation scheme, as the observed data likelihood is virtually impossible to calculate when assuming context-dependent evolution. While a data augmentation approach is not necessary when comparing site-independent models of evolution, we have opted to do so anyway so as to obtain comparable error estimates when calculating (log) Bayes Factors using thermodynamic integration (Lartillot and Philippe 2006). Except when assessing the performance of the model of Pagel and Meade (2004), we use the same MCMC approach as in our previous study. Given the specific structure of the likelihood under the model of Pagel and Meade (2004), we have calculated observed data likelihoods to assess the number of models necessary to maximize model fit as well as to perform model comparisons. Here, we discuss the different

prior distributions we have chosen for the independent and context-dependent evolutionary models tested. The prior distributions $q(\cdot)$ we have assumed in our analyses are listed in Table 1. In this article, we test two versions of the model introduced by Schadt et al. (1998), with the difference being the prior distribution chosen for its eight model parameters. The SLL98 model has eight different instantaneous rates of change, so a first version of the model, denoted SSL98 (8), puts a prior Dirichlet distribution on these eight parameters, requiring them to sum to 1 (before scaling). A second version of the model, denoted SSL98 (12), is inspired by the prior Dirichlet distribution used for the GNR model, where the 12 instantaneous rates of change are required to sum to 1. As the SLL98 (12) model assumes that certain parameters from the GNR model are equal to one another, we hence use a Dirichlet distribution which satisfies the constraint $\alpha + 2\gamma + 2\lambda + 2\delta + \varepsilon + \sigma + \beta + 2\kappa = 1$. We denote the Gamma function with $\Gamma(n)$, with n the number of parameters in the Dirichlet distribution. For example, the Gojobori et al. (1982; GIN82) 6-parameter model in Table 1 has the following distribution properties: $\theta \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1)$, $q(\theta) = \Gamma(6)$ on $0 \leq \theta_i \leq \sum_i \theta_i = 1$. The model of Pagel and Meade (2004) requires estimation of a number of weights w_i that sum to 1. Specifically, under the model of Pagel and Meade (2004), the probability of the data is calculated by summing the likelihood at each site over different matrices $Q_1, Q_2, \dots, Q_j$ and with T the underlying phylogenetic tree, the probability of the data under the mixture model is written as:

Table 1 List of the models tested in this article

Model	P	Prior distribution	Dirichlet parameters
GNR	12	$\theta \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$	a, b, c, d, e, f, g, h, i, j, k, l
GNR-SYM	6	$\theta \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1)$	a, b, c, d, e, f
SSL98 (8)	8	$\theta \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1, 1, 1)$	$\alpha, \gamma, \lambda, \delta, \varepsilon, \sigma, \beta, \kappa$
SSL98 (12)	8	$\theta \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1, 1, 1)$	$\alpha, 2\gamma, 2\lambda, 2\delta, \varepsilon, \sigma, \beta, 2\kappa$
TK81	5	$\theta \sim \text{Dirichlet}(1, 1, 1, 1, 1)$	$\alpha, \beta, \gamma, \delta, \varepsilon$
GIN82	4	$\theta_1 \sim \text{Dirichlet}(1, 1, 1); \theta_2 \sim U(0, 100)$	α, β, γ
GIN82	6	$\theta \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1)$	$\alpha, \beta, \gamma, \delta, \varepsilon, \eta$
TN84	4	$\theta \sim \text{Dirichlet}(1, 1, 1, 1)$	a, b, c, d
B85	4	$\theta \sim \text{Dirichlet}(1, 1, 1, 1)$	a, b, c, d
GTR	6	$\theta \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1)$	$2\pi_A\pi_G\alpha, 2\pi_A\pi_C\beta, 2\pi_A\pi_T\gamma,$ $2\pi_G\pi_C\delta, 2\pi_G\pi_T\varepsilon, 2\pi_C\pi_T\eta$
PM04	6	$\theta \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1)$	$2\pi_A\pi_G\alpha, 2\pi_A\pi_C\beta, 2\pi_A\pi_T\gamma,$ $2\pi_G\pi_C\delta, 2\pi_G\pi_T\varepsilon, 2\pi_C\pi_T\eta$
HG04	12	$\theta \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$	a, b, c, d, e, f, g, h, i, j, k, l

First column model abbreviation, *second column* number of parameters for each model instance, *third column* prior distribution for each model instance, *fourth column* the parametrization of the Dirichlet distribution used to estimate the model parameters. θ represents the instantaneous rates of change, θ_1 and θ_2 are used for the GIN82 4-parameter model as this model contains two sets of parameters which have a different prior distribution and U represents a Uniform distribution

$$P(D|Q_1, Q_2, \dots, Q_j, T) = \prod_i \sum_j w_j P(D_i|Q_j, T),$$

where the summation over j specifies that the likelihood of the data at each site is summed over j separate rate or matrices, the summation being weighted by the weights w_j . In addition to the Dirichlet distribution for the instantaneous rates of change in each of the different matrices, we assume a Dirichlet distribution on the collection of weights as well.

Bayes Factors

To compare the different independent and context-dependent evolutionary models, we have calculated (log) Bayes Factors (Kass and Raftery 1995). Log Bayes Factors are typically divided into four categories depending on their value: from 0 to 1, indicating nothing worth reporting; from 1 to 3, indicating positive evidence of one model over the other; from 3 to 5, indicating strong evidence of one model over the other; and larger than 5, indicating significant (or very strong) evidence of one model over the other. We have chosen to calculate Bayes Factors using thermodynamic integration (Lartillot and Philippe 2006), since the traditional harmonic mean estimator of the marginal likelihood systematically favors parameter-rich models and is hence unfit to compare these complex context-dependent models. We have used the model-switch integration method and have performed bidirectional checks, i.e., we have calculated both annealing and melting integrations under various settings to obtain very similar runs, as suggested in the work of Rodrigue et al. (2006). When comparing different models, we report (log) Bayes Factor estimates for both annealing and melting integrations, as well as their mean. Depending on the dataset analyzed and the model comparisons of interest, we have used different settings for the (log) Bayes Factor calculations to ensure highly accurate results which permit reliable conclusions to be drawn. We will provide more details concerning the number of iterations for each calculation in the appropriate section of this article.

Results

Ancestral Repeats

(Context)-Independent Models

We have tested the independent evolutionary models discussed in the evolutionary models section against the independent GTR model using thermodynamic integration (Lartillot and Philippe 2006). The results for all the independent model comparisons are shown in Table 2, from

Table 2 Ancestral Repeats dataset—context-independent evolutionary models are compared against the independent general time-reversible (GTR) model and ranked according to their performance

Model	Annealing	Melting	logBF
TN84	[−7638.3; −7618.9]	[−7631.6; −7616.5]	−7626.3
GIN82 (6)	[−4377.3; −4349.6]	[−4356.4; −4338.9]	−4355.5
B85	[−110.4; −92.9]	[−100.5; −84.8]	−97.1
SSL98 (12)	[−78.2; −66.5]	[−78.5; −66.7]	−72.4
SSL98 (8)	[−79.5; −67.7]	[−77.8; −65.9]	−72.7
GIN82 (4)	[−73.8; −58.9]	[−65.3; −46.3]	−61.1
TK81	[−69.0; −51.1]	[−51.7; −35.9]	−51.9
GTR	—	—	0
GNR-SYM	[14.5; 36.5]	[35.1; 51.1]	34.3
GNR	[45.0; 60.6]	[56.3; 70.3]	58.1
PM04 (4Q)	[373.9; 379.0]	[388.3; 395.8]	384.2

The general non-reversible model (GNR) offers a significant improvement in model fit over the GTR model while most models show a drastic underperformance relative to the GTR model

which the drastic underperformance of the TN84 and GIN82 6-parameter models is immediately apparent. Even though most of the models listed in Table 2 have six or fewer model parameters, the underperformance of the TN84 and GIN 82 6-parameter models is quite surprising. It can be explained by their specific model structure, of which the substitution rates are shown in Fig. 1. The main substitution patterns that these two models cannot incorporate is the increased substitution from C to T (denoted as rCT) and its compensatory substitution from G to A (denoted as rGA). Indeed, as can be seen from the estimates of the GTR and GNR evolutionary models, these two rates are drastically increased compared to the other rates. While the GIN82 6-parameter model does require the rCT and rGA parameters to be equal, it also requires these two parameters to be equal to both the rGT and the rCA parameters, which causes an artificial decrease in the estimated values of the rCT and rGA parameters. The TN84 model does not require rCT and rGA to be equal but requires that rCT equals both rAT and rGT, while rGA is forced to be equal to rCA and rTA. For both the models, a similar reasoning holds for the rAC and rTG parameters, which cannot be estimated accurately.

The GNR model significantly outperforms the GTR model in model fit to the Ancestral Repeats dataset. The log Bayes Factor versus the GTR model shows a difference in model fit that is nowhere near the difference between the GTR model and the TN84 or GIN82 6-parameter model. The fact that the GNR model significantly outperforms the GTR model can be attributed to flexibility issues within the GTR model, where each model (or matrix) entry is a product of a base frequency and a model parameter. In all, the GTR model has six exchangeability parameters to be estimated, but also

Fig. 1 Graphical explanation of the underperformance of the TN84 and GIN82 6-parameter models and the outperformance of the GNR (12 parameters) model vis-à-vis the GTR model. Shown are means and corresponding 95% credibility intervals instantaneous rates of change from 100,000 MCMC iterations, discarding the first 20,000 as burn-in

has four equilibrium frequency parameters. This gives it less flexibility than the GNR model, its 12 parameters being justified by the model comparison outcome.

The two SSL98 models result in a very similar decrease in model fit when compared to the independent GTR model, based on their corresponding confidence intervals for the log Bayes Factors. This suggests that the choice of prior distribution for the SSL98 model had no influence on the decrease in model fit when compared to the independent GTR model. We have tested this by calculating Bayesian *P* values (e.g., Schafer 1997) from 80,000 posterior estimates (having discarded the first 20,000 as burn-in). None of the 8 model parameters could be shown to differ significantly at the 5% level

Before being able to calculate a difference in model fit between the independent PM04 and GTR models, we must first assess how many rate matrices are needed in the PM04 model. Indeed, as mentioned in the work of Pagel and Meade (2004), the number of rate matrices is not usually known in advance and once the main patterns are modelled using a certain number of rate matrices, further adding rate matrices will not improve the likelihood (which is analogous to determining how many rate categories to estimate under a gamma model for among-site rate variation). Figure 2 shows the average log-likelihoods (from 80,000

Fig. 2 Plot of the average log-likelihoods of the model of Pagel and Meade (2004) using a range of rate matrices for the Ancestral Repeats dataset. Average log-likelihoods were calculated from 100,000 iterations, discarding the first 20,000 as burn-in. The positive trend of the curve connecting average log-likelihoods diminishes after four rate matrices, giving rise to our decision of using four rate matrices to perform a model comparison with the independent GTR model

log-likelihood estimates, having discarded the first 20,000 as burn-in) for the Ancestral Repeats dataset with an increasing number of rate matrices. Given that (log) Bayes Factor calculation using thermodynamic integration is computationally demanding, one has to be pragmatic in the number of rate matrices to be used in a model comparison. As the major increases in average log-likelihood occur up to four rate matrices, we only perform a model comparison for this scenario and we denote this model as PM04 (4Q) in Table 2. The bidirectional mean log Bayes Factor versus the independent GTR model equals 384.2 log units, drastically outperforming all the other context-independent models tested. Note that we have previously shown that the varying rates across sites assumption using $n = 4$ discrete rate classes (Yang 1996) results in a mean log Bayes Factor of 355.6 (annealing: [334.8; 348.4]; melting: [362.5; 376.7]) when compared to the independent GTR model, and that further increases in the number of rate categories did not result in additional increases in model fit (Baele et al. 2009). This means that the model of Pagel and Meade (2004) results in a higher increase in model fit than the assumption of varying rates across sites, in accordance with the results presented by Pagel and Meade (2004) in, e.g., their dataset of 12S ribosomal RNA.

Context-Dependent Models

For each independence model listed in Table 1, we develop a context-dependent version for which we use the same notation as their independent counterparts. We have previously developed two context-dependent models based on the GTR model: a first model with a set of independent base frequencies (GTR16C; Baele et al. 2008) and a second model with context-dependent equilibrium frequencies (GTR-CDEF; Baele et al. 2010). We also include the

model developed by Hwang and Green (2004) in our analysis. For all these models, we calculate the log Bayes Factor against the independent GTR model.

As we have previously shown (Baele et al. 2010), a second-order Markov chain is required at the ancestral root sequence to accurately infer the context-dependent model parameters and to optimize the parameter-performance trade-off. Such a second-order Markov chain requires 64 parameters to be estimated, which we include in each model's parameter count. Given our previous finding that a third-order Markov chain at the ancestral root sequence does not further improve the model fit to the data and that model comparisons for the Ancestral Repeats dataset is computationally demanding, we have refrained from calculating such third-order Markov chains here (Baele et al. 2010).

Table 3 shows that the context-dependent model of Hwang and Green (2004) outperforms all other context-dependent models discussed in this article, in terms of increase in model fit when compared to the independent GTR model, including the most parameter-rich model based on the GNR model. Both these context-dependent models outperform our previously introduced context-dependent model with context-dependent equilibrium frequencies (Baele et al. 2010). We have tested, by calculating Bayesian P values (e.g., Schafer, 1997) based on 100,000 iterations—discarding the first 20,000 as burn-in—of our MCMC approach, whether the increase in model fit of the model of Hwang and Green (2004) solely results from the decrease in number of parameters when compared to the context-dependent GNR model, or if it is the result of different (and hence better suited) context-dependent parameter estimates. Only 2 out of the 192 parameter estimates (1 set of 192 unique parameters for the context-dependent GNR model and 2 sets of 96 identical parameters for the HG04 model) were significantly different at the

Table 3 Ancestral Repeats dataset—context-dependent evolutionary models are compared against the independent general time-reversible (GTR) model and ranked according to their performance

While the TN84 model (like its independent version) performs significantly worse than the GTR model, all the other context-dependent models significantly outperform the GTR model. The largest increase in model fit is obtained for the model of Hwang and Green (2004; HG04)

Model	Param.	Annealing	Melting	logBF
TN84	128	[−2202.4; −2166.6]	[−2175.1; −2123.5]	−2166.9
GTR	6	–	–	0
GIN82 (6)	160	[854.7; 896.8]	[891.6; 933.8]	894.2
GIN82 (4)	128	[5133.2; 5170.2]	[5190.8; 5228.7]	5180.7
B85	128	[5156.7; 5193.0]	[5156.7; 5193.0]	5190.0
TK81	144	[5201.7; 5238.3]	[5239.7; 5303.1]	5245.7
GNR-SYM	160	[5298.6; 5340.5]	[5334.0; 5377.1]	5337.5
GTR16C	160	[5573.4; 5608.4]	[5593.5; 5630.3]	5601.4
SSL98 (8)	192	[7076.5; 7108.5]	[7085.7; 7118.1]	7097.2
SSL98 (12)	192	[7105.5; 7137.2]	[7109.6; 7141.1]	7123.3
GTR-CDEF	160	[7181.4; 7228.1]	[7225.9; 7277.1]	7228.1
GNR	256	[7309.7; 7351.8]	[7338.9; 7384.6]	7346.3
HG04	160	[7481.3; 7525.3]	[7511.5; 7556.2]	7518.6

Fig. 3 Context-dependent evolutionary model parameter estimates for the model of Hwang and Green (2004) for the Ancestral Repeats dataset. Shown are means with corresponding 95% credibility intervals

5% level, before applying a multiple testing correction: the rTC parameter in the AXG context (Bayesian P value: 0.03) and the rTA parameter in the CXT context (Bayesian P value: 0.04). Applying a simple Bonferroni correction for multiple testing will hence result in no significantly different parameter estimates between both the models.

Figure 3 shows the parameters estimates for the HG04 model. The main difference between these estimates and those from the context-dependent GNR model lies in the decreased widths for the 95% credibility intervals for the parameters (data not shown), due to the fact that each parameter is used to describe a substitution type at opposing sides of a stem region, resulting in a more drastic change in likelihood when proposing new parameter values. Figure 3 clearly illustrates the presence of the CpG-methylation-deamination process in mammals. The mean substitution rates from C to T (rCT) in the AXG, CXG,

GXG, and TXG contexts (and hence also the mean substitution rates from G to A (rAG) in the CXT, CXG, and CXA contexts) are 12.69, 11.79, 8.07, and 10.34, respectively. Such compensatory mutations make sense in this dataset as three of the nine genes of the dataset we have analyzed are located on the opposite strand from the other six as discussed by Green et al. (2003) and Siepel and Haussler (2004).

Four of the twelve context-dependent models listed in Table 3 result in a log Bayes Factor between 5,000 and 5,500 log units when compared to the independent GTR model: GIN82 (4-parameter version), B85, TK81, and GNR-SYM. The reason for this increase in performance is solely due to the fact that these four models assume that in each context the rCT and rGA parameter values are equal. This prohibits these context-dependent models to correctly estimate the CpG-effects in all but one context: the CXG context.

The two context-dependent SSL98 models result in a (relatively small) difference in model fit when compared to the independent GTR model. As discussed earlier, we have invested additional computational efforts to ensure that both annealing and melting confidence intervals overlap sufficiently. The choice of prior distribution for the SSL98 model thus seems to have an influence on the performance of the SSL98 model. Both versions yield highly similar parameter estimates, which we have tested by calculating Bayesian P values (e.g., Schafer, 1997). None of the 128 model parameters could be shown to differ significantly at the 5% level. The difference in increase in model fit between both the models stems from a difference in performance when no data is available for the context-dependent models (i.e., when the thermodynamic integration scheme finds itself in the prior of the context-dependent model; data not shown). In other words, the prior of the SSL98 (12) model describes the context-dependent evolutionary parameters better than the prior of the SSL98 (8) model.

Parameter Reduction

We have tested whether the context-dependent GNR model would benefit from clustering similar evolutionary parameters across contexts, i.e., clustering those parameters that have a similar distribution in all 16 contexts, as this model has clearly the largest number of evolutionary parameters. To determine which evolutionary parameters would be best suited for parameter reduction, we have applied a 16-dimensional principal component clustering for the 12 substitution types in the context-dependent GNR model (see Fig. 4). The different evolutionary parameters are of different degrees of magnitude, with those parameters related to the CpG-effects (i.e., rCT and rGA) clearly the largest ones, which is why Fig. 4 shows four clusters (a)–(d), progressively omitting the largest parameters. The most promising candidates for clustering parameters across contexts are hence: rAC, rAT, rTA, and rTG. The different clustering steps and their impact on the model's performance are shown in Table 4. As can be seen from Table 4, two models compete for the best parameter-performance trade-off: (1) the model that only clusters rAC with rAT, with 240 parameters and (2) the model that clusters rAC with rAT and rTA with rTG, with 224 parameters. Further clustering only decreases the model fit, hence it can be concluded that a simple across-contexts parameter clustering approach does not drastically increase model fit for the Ancestral Repeats dataset. Given that the HG04 model gives rise to a similar clustering (but with clusters even further apart from one another; data not shown), we did not attempt to perform this clustering approach for the HG04 model as we do not expect large increases in model fit

relative to the magnitude of the log Bayes Factor for the HG04 model.

Pseudogenes

(Context)-Independent Models

We have performed the same set of model comparisons for the Ancestral Repeats dataset on our pseudogenes dataset. The results for all the independent model comparisons for the pseudogenes dataset are shown in Table 5. Compared to the Ancestral Repeats dataset, we can conclude that the GNR-SYM model again significantly improves the model fit, whereas the GNR model is not able to yield a significant difference in model fit when compared to the independent GTR model. Again, the TN84 and the GIN82 6-parameter models show a drastic underperformance when compared to the GTR model. The two SSL98 models result in a very similar decrease in model fit when compared to the independent GTR model, based on the confidence intervals for the corresponding log Bayes Factors. In other words, the choice of prior distribution for the SSL98 model had no influence on the decrease in model fit when compared to the independent GTR model, as was the case for the Ancestral Repeats dataset.

As for the Ancestral Repeats dataset, we must first assess how many rate matrices are needed in the PM04 model before being able to calculate a difference in model fit between the independent PM04 and GTR models. Figure 5 shows the average log-likelihoods (from 80,000 log-likelihood estimates, having discarded the first 20,000 as burn-in) for the Pseudogenes dataset with an increasing number of rate matrices. Only a minor increase in average log-likelihood occurs for the case of two rate matrices, for which we perform a model comparison and denote this model as PM04 (2Q) in Table 5. Based on the bidirectional mean log Bayes Factor, we conclude that for the Pseudogenes dataset, the model of Pagel and Meade (2004) is unable to provide a significant increase in model fit, compared to the independent GTR model. Indeed, a further increase in the number of rate matrices does not further increase the average log-likelihood with an increasing number of parameters. For a larger number of rate matrices, the log Bayes Factors will only decrease.

The assumption of varying rates across sites using $n = 4$ discrete rate classes results in a slightly larger increase in model fit with a bidirectional mean log Bayes Factor of 5.4 units (annealing: [5.5; 5.7]; melting: [5.0; 5.2]), corresponding to the findings of Yang (1994). For the Pseudogenes dataset, it can hence be concluded that the model of Pagel and Meade (2004) offers approximately the same increase in model fit as the assumption of varying rates across sites.

Fig. 4 Sixteen-dimensional principal components clustering of the 12 different substitution rates from the GNR12P16C model (i.e., those substitution rates clustered closely together have similar estimates across all 16 neighboring base combinations). From this clustering,

the following clustering attempts to reduce dimensionality can be proposed (in descending order): **(a)** rAC with rAT; **(b)** rTA with rTG; **(c)** rAC with rAT and rTA with rTG; **(d)** rAC with rAT, rTA, rTG

Context-Dependent Models

Given the smaller size (and hence the lower computational demands for performing model comparisons) of the pseudogenes dataset when compared to the Ancestral Repeats dataset, we have calculated log Bayes Factors for zero-, first-, second- and third-order Markov chains at the ancestral root sequence. All context-dependent models compared showed the largest increase in model fit when assuming a first-order Markov chain at the ancestral root, followed by a second-order Markov chain (with a difference of about 20 log units in terms of model fit; data not shown). Such a first-order Markov chain requires 16 parameters to be estimated, which we include in each model's parameter count. The results for all the context-dependent model comparisons for the pseudogenes dataset are shown in Table 6.

Table 6 shows that, for the pseudogenes dataset as well, the context-dependent model of Hwang and Green (2004) outperforms all other context-dependent models discussed in this article. To find such dependency patterns in such a (relatively) small dataset, with such strong support, provides evidence that even small datasets may benefit from parameter-rich context-dependent models, which are often evaluated on large genomic datasets.

Figure 6 shows the parameter estimates for the HG04 model for the pseudogenes dataset. As is the case for the Ancestral Repeats dataset, the presence of the CpG-methylation-deamination process can easily be established for the pseudogenes dataset. The mean substitution rates from C to T (rCT) in the AXG, CXG, GXG, and TXG contexts (and hence also the mean substitution rates from G to A (rAG) in the CXT, CXG, CXC, and CXG contexts) are

Table 4 Ancestral Repeats dataset—across-contexts parameter clustering (reducing parameter space with 16 evolutionary parameters at a time), starting from the context-dependent GNR model, only results in a marginal increase in terms of model fit for the Ancestral Repeats dataset

Clustering step	Param.	Annealing	Melting	logBF
GNR	256	[7309.7; 7351.8]	[7338.9; 7384.6]	7346.3
1	240	[7330.8; 7373.2]	[7375.2; 7421.1]	7375.1
2	240	[7319.4; 7361.9]	[7345.8; 7391.8]	7354.7
3	224	[7341.1; 7385.8]	[7364.3; 7411.8]	7375.8
4	208	[7311.2; 7353.5]	[7338.0; 7380.6]	7345.8

The clustering steps, as proposed from Fig. 4 are: 1. rAC with rAT; 2. rTA with rTG; 3. rAC with rAT and rTA with rTG, and 4. rAC with rAT, rTA, rTG

Table 5 Pseudogenes dataset—context-independent evolutionary models are compared against the independent general time-reversible (GTR) model and ranked according to their performance

Model	Annealing	Melting	logBF
TN84	[−307.6; −295.1]	[−303.3; −294.5]	−300.1
GIN82 (6)	[−169.2; −162.9]	[−169.2; −159.9]	−165.3
SSL98 (8)	[−9.4; −5.7]	[−8.7; −5.2]	−7.3
SSL98 (12)	[−7.3; −3.6]	[−7.4; −3.6]	−5.5
GNR	[−2.5; −0.0]	[−4.1; 0.4]	−1.6
B85	[−26.8; 15.2]	[−1.7; 8.7]	−1.1
GTR	—	—	0
TK81	[−6.3; 3.5]	[−0.0; 11.1]	2.1
PM04 (2Q)	[0.5; 4.0]	[2.1; 8.7]	3.8
GIN82 (4)	[1.9; 8.0]	[3.6; 12.8]	6.6
GNR-SYM	[9.5; 18.6]	[9.1; 18.2]	13.9

The general non-reversible symmetrical model (GNR-SYM) offers a significant improvement in terms of model fit over the GTR model

5.85, 7.24, 3.86, and 4.67, respectively. Compared to the Anstral Repeats dataset, these substitution rates are less elevated, but the presence of the CpG-effect in this dataset is unmistakable. The uncertainty of these estimates (shown in Fig. 6 by the 95% credibility intervals) is much higher due to less data being available to estimate the parameters.

Parameter Reduction

Given the drastic increase in uncertainty for the context-dependent model parameter estimates, the pseudogenes dataset may greatly benefit from the suggested clustering approach. Further, the small(er) size of this dataset enables calculation of more MCMC iterations, allowing for more model comparisons to be made. For the pseudogenes dataset, we have thus tested whether the context-dependent GNR model and the HG04 model benefit from parameter clustering. Figure 7 shows the 16-dimensional principal

Fig. 5 Plot of the average log-likelihoods of the model of Pagel and Meade (2004) using a range of rate matrices for the Pseudogenes dataset. Average log-likelihoods were calculated from 100,000 iterations, discarding the first 20,000 as burn-in. The curve connecting the average log-likelihoods does not show any drastic increase, leading to the assumption that using 2 or more rate matrices will not lead to significant (log) Bayes Factor in the model comparison with the general time-reversible model

Table 6 Pseudogenes dataset—context-dependent evolutionary models are compared against the independent general time-reversible (GTR) model and ranked according to their performance

Model	Param.	Annealing	Melting	logBF
TN84	80	[−148.5; −135.3]	[−142.1; −126.8]	−138.2
GIN82 (6)	112	[−19.8; −7.6]	[−22.5; −8.0]	−14.5
GTR	6	—	—	0
GNR	208	[118.0; 124.0]	[119.4; 125.5]	121.7
SSL98 (8)	144	[124.4; 130.3]	[125.8; 132.2]	128.2
GTR16C	112	[127.0; 137.1]	[125.6; 138.0]	131.9
B85	80	[126.3; 142.3]	[140.3; 154.9]	140.9
GNR-SYM	112	[131.8; 145.7]	[137.8; 153.1]	142.1
TK81	96	[131.6; 144.5]	[138.0; 155.0]	142.3
GIN82 (4)	80	[139.5; 152.2]	[134.1; 152.7]	144.6
SSL98 (12)	144	[149.1; 155.0]	[148.2; 153.9]	151.5
GTR-CDEF	112	[152.5; 160.0]	[152.4; 159.7]	156.2
HG04	112	[168.9; 174.9]	[171.4; 177.3]	173.1

While the TN84 model (like its independent version) performs significantly worse than the GTR model, all but one of the other context-dependent models significantly outperform the GTR model. The largest increase in model fit is again obtained for the model of Hwang and Green (2004; HG04)

component clustering for the 12 substitution types in the HG04 model (similar for the context-dependent GNR model; data not shown).

Fig. 6 Context-dependent evolutionary model parameter estimates for the model of Hwang and Green (2004) for the pseudogenes dataset. Shown are means with corresponding 95% credibility intervals

The different clustering steps as well as the resulting increase in model fit compared to the independent GTR model can be seen in Tables 7 (for the context-dependent GNR model) and 8 (for the HG04 model). Table 7 shows that the proposed clustering approach is able to reduce the model's parameter count by 80 parameters by successively clustering highly similar parameters across contexts. Based on the bidirectional log Bayes Factor means, this decrease in number of parameters yields an increase in model fit of approximately 40 log units when compared to the independent GTR model.

Even after performing these clustering steps, the optimized context-dependent GNR model is unable to outperform the HG04 model. Further, Table 8 shows that the HG04 model can be further optimized by clustering parameters. Even though the HG04 was already the best

performing context-dependent model, an additional increase in model fit of approximately 30 log units was obtained by the cluster approach, reducing the model's parameter count by 40 parameters. The clustering of the different parameters (see Fig. 6) yields similar clustering scenarios for both the context-dependent GNR model and the HG04 model. The optimal model in both the cases also clusters the same parameters together, i.e., the final model in both Tables 7 and 8 has the same parameters clustered together.

Discussion

In this article, we have presented a series of context-dependent evolutionary models, with different structures

Fig. 7 Sixteen-dimensional principal components clustering of the 12 different substitution rates from the HG04 model (i.e., those substitution rates clustered closely together have similar estimates

aimed at incorporating complex substitution patterns and consisting of different numbers of parameters. We have compared the different context-dependent models by calculating log Bayes Factors using thermodynamic integration (Lartillot and Philippe 2006) and found that for both the datasets, the context-dependent model of Hwang and Green (2004) yielded the largest increase in model fit when compared to the independent GTR model. Unlike Hwang and Green (2004), we have assumed a lineage-independent context-dependent evolutionary model for both the datasets. For the two datasets, which consist of closely related sequences, this assumption is in accordance with the analysis of Hwang and Green (2004), who also used a single model for the clade of primates.

Mixture models of evolution, such as the model of Pagel and Meade (2004), may yield large increases in model fit when compared to the GTR model, while enabling

across all 16 neighboring base combinations). The clustering scenario for the context-dependent GNR model is slightly different (data not shown)

phylogenetic tree reconstruction through their assumption of site-independent evolution. However, as demonstrated in this article, site-independent models are not able to capture context-dependent patterns of evolution, such as the so-called CpG-effect. This requires conditioning the evolution of a given site on the identities of its neighboring bases. Mixture models nonetheless provide food for thought as they raise the question whether a mixture context-dependent model would make sense and whether such a combination would further improve model fit. For example, for (ancestral) sites where a C has a G as its 3' neighbor, a mixture context-dependent model with several rate classes would be able to entertain the possibility that a model containing a high rCT substitution rate is applied to these sites with a large probability, while the remaining models, possibly with much lower rCT rates, are applied to these sites with a low probability.

Table 7 Pseudogenes dataset—across-contexts parameter clustering (reducing parameter space with 16 evolutionary parameters at a time), starting from the context-dependent GNR model, results in a significant increase in terms of model fit for the pseudogenes dataset

Clustering step	Param.	Annealing	Melting	logBF
GNR	256	[118.0; 124.0]	[119.4; 125.5]	121.7
1	240	[127.4; 133.5]	[129.3; 135.3]	131.3
2	240	[124.7; 130.7]	[125.7; 131.9]	128.2
3	208	[150.1; 156.0]	[150.5; 156.3]	153.2
4	192	[148.9; 154.5]	[147.1; 153.0]	150.9
5	192	[155.5; 161.5]	[155.4; 161.2]	158.4
6	192	[156.2; 161.9]	[157.7; 163.6]	159.8
7	176	[160.7; 166.5]	[161.9; 167.6]	164.2

The clustering steps, as proposed from Fig. 6 are: 1. rAC with rAT; 2. rTA with rTG; 3. rAC with rAT and rTA with rTG; 4. step 3 and rAG with rTC; 5. step 3 and rCA with rCG; 6. step 3 and rGC with rGT; and 7. step 3 and rCA with rCG and rGC with rGT. Further clustering steps did not result in additional gains in model fit

Table 8 Pseudogenes dataset—across-contexts parameter clustering (reducing parameter space with 16 evolutionary parameters at a time), starting from the HG04 model, results in a significant increase in terms of model fit for the pseudogenes dataset

Clustering step	Param.	Annealing	Melting	logBF
HG04	160	[168.9; 174.9]	[171.4; 177.3]	173.1
1	152	[175.0; 180.9]	[174.4; 180.3]	177.6
2	152	[177.1; 183.0]	[176.1; 181.9]	179.5
3	136	[193.0; 198.7]	[194.5; 200.4]	196.6
4	120	[198.5; 204.0]	[202.2; 208.5]	203.3

The clustering steps for the HG04 model are: 1. rAT with rTA; 2. rAC with rTG; 3. rAT with rTA and rAC with rTG; and 4. step 3 and rCA with rCG and rGC with rGT. Further clustering steps did not result in additional gains in model fit

We have shown that modeling context-dependent substitution patterns may result in increases in model fit, both in a large Ancestral Repeats dataset and in a small Pseudogenes dataset, that are several times as large than the increase brought about by assuming varying rates across sites. Given that the assumption of varying rates across sites has been most beneficial in the field of phylogenetics (e.g., Yang 1996), we anticipate that modeling site-specific dependencies may have an important impact on several aspects of phylogenetic inference. For example, clade posterior probabilities and even the tree topology may be estimated more accurately using these context-dependent models. However, using context-dependent models in inferring phylogenetic trees is still an issue (e.g., Ronquist and Deans 2010).

The fact that modeling context-dependence yields large increases in model fit in primate non-coding sequences, especially when compared to the assumption of varying

rates across sites, could mean that many datasets of various types of organisms could benefit from the model presented here. Primate and mammalian datasets are the first choice to test such models because of the CpG-methylation-deamination process known to exist in mammalian evolution. However, the inspiration for our study came from findings of Morton et al. (1997) who reported an influence of neighboring bases on substitution bias in non-coding regions of the plant chloroplast genome. In other words, non-coding plant datasets may exhibit context-dependent substitution patterns which may in turn influence phylogenetic tree reconstruction in the sense that previously unresolved branching patterns may be solved when the context-dependent models offer significant increases in model fit.

Performing model comparisons using the thermodynamic integration approach to calculate Bayes Factors (Lartillot and Philippe 2006) is computationally demanding, so to make this approach feasible, we have split each log Bayes Factor calculation over multiple processor cores (Baele et al. 2009). For both the datasets, the calculation settings (number of cores, MCMC iterations and days of computation should the calculations occur on a single processor core) are shown in Table 9. Note that for both the context-independent and the context-dependent versions of the SSL98 model, more accurate settings were used as we were interested in accurately assessing the influence of the prior distribution on the outcome of the model comparison with the GTR model.

Table 9 Settings for the thermodynamic integration approach to calculate Bayes Factors for both the Ancestral Repeats and Pseudogenes datasets

Model(s)	Cores	MCMC iterations	Days of computation
Ancestral Repeats dataset			
SSL98 (CI)	40	3,680,000	644
Other models (CI)	40	2,240,000	392
SSL98 (CD)	40	5,440,000	1.088
Other models (CD)	40	800,000	186
Pseudogenes dataset			
SSL98 (CI)	1	22,400,000	64
Other models (CI)	1	2,000,000	8
TN84, GIN82 (6), GTR, B85, TK81 (CD)	1	2,000,000	15
Other models (CD)	1	14,400,000	111

The *first column* shows the evolutionary model (CI context-independent, CD context-dependent); the *second column* reports the number of processor cores used in the computation; the *third column* shows the number of MCMC iterations, not including the burn-in, performed; and the *fourth column* reports the computation time should the calculations be done on a single processor core

As can be seen from the clustering efforts for the pseudogenes dataset, an optimized version of the context-dependent GNR model is unable to outperform the model of Hwang and Green (2004). This is due to the specific clustering approach and illustrates a crucial aspect of model building, i.e., that “better, more realistic models should not mean more parameter-rich models” (Steel 2005). In other words, the model of Hwang and Green (2004) takes full advantage of biologically realistic assumptions by providing a single set of parameters for compensatory mutations, thereby incorporating structural information in the design of their model.

While the context-dependent model of Hwang and Green (2004) leads to the largest increase in model fit for both the datasets, we have shown that the model could be further refined to increase model fit for the pseudogenes dataset. Even though the model of Hwang and Green (2004) already assumes that compensatory mutations occur on opposing regions, it is impossible to predict if such a model can be further reduced by clustering similar parameters. The process of repeatedly assessing model performance and clustering parameters is computationally intensive. An automated approach to this problem, e.g., using Reversible Jump MCMC (e.g., Green, 1995), therefore seems preferable. However, such an approach leads to additional programming, may require longer run times (i.e., more MCMC iterations) and may exhibit poor convergence.

Each of the two primate datasets requires a different Markov chain at the ancestral root sequence, for which the main reason seems to be the amount of data available in the dataset. For the large Ancestral Repeats dataset, we had already established strong support for the use of a second-order Markov chain at the ancestral root (Baele et al. 2010). We have shown that the smaller pseudogenes dataset only requires a first-order Markov chain at the ancestral root. While the clustering approach we performed in this article is solely aimed at reducing the number of parameters in the context-dependent evolutionary model itself, the number of parameters of the Markov chain describing the ancestral root sequence can most likely also be reduced. In other words, the Ancestral Repeats dataset may benefit from using a third-order Markov chain at the root, with most of the parameters clustered together, which would lead to a fully parameterized second-order Markov chain at the root with a few specific third-order dependencies. Similarly, the ancestral root distribution of the pseudogenes dataset could consist of a heavily reduced second-order Markov chain, leading to a fully parameterized first-order Markov chain at the root with a few specific second-order dependencies. Modeling such reduced Markov chains could be accommodated using a Reversible Jump MCMC approach (Green 1995).

In this article, we have discussed the importance of allowing substitution patterns at a given site to depend

upon the identities of that site’s two immediate flanking neighbors. It is however unlikely that only these two neighbors are responsible for specific context-dependent patterns of evolution. Allowing for longer-range dependencies will address interesting topics such as whether CpG-effects are in fact only dependent upon guanine as the 3′ neighbor or whether other dependencies influence the accompanying substitution rates. The extension of the presented approach to model the influence of more than two neighboring sites is the subject of ongoing research.

Acknowledgments Computational resources and services used in this study were provided by Ghent University. We would like to thank Stijn De Weirdt in particular for his assistance in using Ghent University’s High Performance Computing facility. We would also like to thank Hervé Philippe and one anonymous referee for providing useful comments on a first version. Yves Van de Peer acknowledges support from an Interuniversity Attraction Pole (IUAP) grant for the BioMaGNet project (Bioinformatics and Modelling: from Genomes to Networks, ref. p6/25). Stijn Vansteelandt acknowledges support from IAP research network Grant No. P06/03 from the Belgian government (Belgian Science Policy). We acknowledge the support of Ghent University (Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks”).

References

- Baele G, Van de Peer Y, Vansteelandt S (2008) A model-based approach to study nearest-neighbor influences reveals complex substitution patterns in non-coding sequences. *Syst Biol* 57(5):675–692
- Baele G, Van de Peer Y, Vansteelandt S (2009) Efficient context-dependent model building based on clustering posterior distributions for non-coding sequences. *BMC Evol Biol* 9:87
- Baele G, Van de Peer Y, Vansteelandt S (2010) Modelling the ancestral sequence distribution and equilibrium frequencies in context-dependent models for primate non-coding sequences. *BMC Evol Biol* (submitted, under review)
- Balakirev ES, Ayala FJ (2003) Pseudogenes: are they “junk” or functional DNA? *Annu Rev Genet* 37:123–151
- Blaisdell BE (1985) A method for estimating from two aligned present day DNA sequences their ancestral composition and subsequent rates of composition and subsequent rates of substitution, possibly different in the two lineages, corrected for multiple and parallel substitutions at the same site. *J Mol Evol* 22:69–81
- Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W (2004) Aligning multiple genomic sequences with the Threaded Blockset Aligner. *Genome Res* 14:708–715
- Fryxell KJ, Zuckerkandl E (2000) Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol* 17(9):1371–1383
- Gojbori T, Ishii K, Nei M (1982) Estimation of the average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *J Mol Evol* 18:414–423
- Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711–732
- Green P, Ewing B, Miller W, Thomas PJ, NISC Comparative Sequencing Program, Green ED (2003) Transcription-associated

- mutational asymmetry in mammalian evolution. *Nat Genet* 33:514–517
- Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22:160–174
- Holmquist R (1976) Solution to a gene divergence problem under arbitrary stable nucleotide transition probabilities. *J Mol Evol* 8:337–349
- Huelsenbeck JP, Bollback JP, Levine AM (2002) Inferring the root of a phylogenetic tree. *Syst Biol* 51(1):32–43
- Hwang DG, Green P (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA* 101:13994–14001
- Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro MN (ed) *Mammalian protein metabolism*, vol III. Academic Press, New York, pp 21–132
- Jurka J (2000) RepBase Update: a database and an electronic journal of repetitive elements. *Trends Genet* 9:418–420
- Kass RE, Raftery AE (1995) Bayes Factors. *J Am Stat Assoc* 90(430):773–795
- Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16:111–120
- Kimura M (1981) Estimation of evolutionary differences between homologous nucleotide sequences. *Proc Natl Acad Sci USA* 78:454–458
- Lanavé C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. *J Mol Evol* 20:86–93
- Lartillot N, Philippe H (2006) Computing Bayes Factors using thermodynamic integration. *Syst Biol* 55(2):195–207
- Margulies EH, Blanchette M, NISC Comparative Sequencing Program, Haussler D, Green ED (2003) Identification and characterization of multi-species conserved sequences. *Genome Res* 13:2507–2518
- Margulies EH, Chen CW, Green ED (2006) Differences between pairwise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends Genet* 22(4):187–193
- Mighell AJ, Smith NR, Robinson PA, Markham AF (2000) Vertebrate pseudogenes. *FEBS Lett* 468:109–114
- Miyamoto MM, Slighton JL, Goodman M (1987) Phylogenetic relations of humans and African apes from DNA sequences in the $\psi\eta$ -globin region. *Science* 238:369–373
- Morton BR, Oberholzer VM, Clegg MT (1997) The influence of specific neighboring bases on substitution bias in noncoding regions of the plant chloroplast genome. *J Mol Evol* 45:227–231
- Pagel M, Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol* 53(4):571–581
- Rodrigue N, Philippe H, Lartillot N (2006) Assessing site-interdependent phylogenetic models of sequence evolution. *Mol Biol Evol* 23(9):1762–1775
- Ronquist F, Deans AR (2010) Bayesian phylogenetics and its influence on insect systematic. *Annu Rev Entomol* 55:189–206
- Schadt EE, Sinsheimer JS, Lange K (1998) Computational advances in maximum likelihood methods for molecular phylogeny. *Genome Res* 8:222–233
- Schaffer JL (1997) *Analysis of incomplete multivariate data*. Chapman & Hall/CRC, London
- Siepel A, Haussler D (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* 21(3):468–488
- Smit AFA, Hubley R, Green P (1996–2004) RepeatMasker Open-3.0. <http://www.repeatmasker.org>
- Steel M (2005) Should phylogenetic models be trying to ‘fit an elephant’? *Trends Genet* 21(6):307–309
- Tajima F, Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. *Mol Biol Evol* 1(3):269–285
- Takahata N, Kimura M (1981) A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics* 98:641–657
- Tamura K, Nei M (1984) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10(3):512–526
- Yang Z (1994) Estimating the pattern of nucleotide substitution. *J Mol Evol* 39:105–111
- Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* 11(9):367–372
- Zwickl D, Holder M (2004) Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. *Syst Biol* 53(6):877–888