A nucleotide substitution model with nearest-neighbour interactions

Gerton Lunter and Jotun Hein

Bioinformatics group, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: It is well known that neighbouring nucleotides in DNA sequences do not mutate independently of each other. In this paper, we introduce a context-dependent substitution model and derive an algorithm to calculate the likelihood of sequences evolving under this model. We use this algorithm to estimate neighbour-dependent substitution rates, as well as rates for dinucleotide substitutions, using a Bayesian sampling procedure. The model is irreversible, giving an arrow to time, and allowing the position of the root between a pair of sequences to be inferred without using out-groups.

Results: We applied the model upon aligned human–mouse non-coding data. Clear neighbour dependencies were observed, including 17–18-fold increased $\text{CpG}$ to $\text{TpG}$/$\text{CpA}$ rates compared with other substitutions. Root inference positioned the root halfway the mouse and human tips, suggesting an approximately clock-like behaviour of the irreversible part of the substitution process.

Contact: lunter@stats.ox.ac.uk

INTRODUCTION

Most current stochastic models for the evolutionary nucleotide substitution process in DNA sequences assume that neighbouring sites evolve independently. This considerably simplifies calculations, since under this assumption, the likelihood is the product of individual site likelihoods. However, there is ample evidence that independence is violated (Karlin and Burge, 1995), e.g. by the well-known methylation-induced rate increase of $\text{C}$ to $\text{T}$ (and $\text{G}$ to $\text{A}$) substitutions in vertebrate $\text{CpG}$ dinucleotides. The importance of neighbour dependencies in the substitution process has long been recognized, and several ways of modelling these dependencies have been proposed. For example, Siepel and Haussler (2003) show that a Markov chain along a pair of sequences fits sequence data substantially better than a series of independent pairwise nucleotide distributions (a ‘zeroth-order’ Markov chain) do.

A natural model for context-dependent substitutions is the one that assigns rates to all possible dinucleotide-to-dinucleotide substitution, which then apply to all overlapping neighbouring nucleotide pairs in a sequence. This model, referred to as the ‘dinucleotide model’, is arguably the simplest possible general evolutionary model that takes neighbour dependencies into account, and captures (neighbour-independent) dinucleotide substitutions as well. One of its essential features is that long-range dependencies between sites immediately arise, due to the possibility of overlapping hits. Because of this ‘contagious dependence’, this model is harder to analyse than independent-site models. Previous studies by Jensen and Pedersen (2000) and Arndt et al. (2003) used similar explicit evolutionary models, both in the nucleotide and codon contexts, and focussed primarily on the relation between the rate matrix and the equilibrium sequence distribution. As the equilibrium distribution contains only partial information on the substitution rates, this can be used only for estimation of sparsely parameterized models. For richer models, it is necessary to have a method for calculating the likelihood of observing a pair of homologous sequences. For richer models, it is necessary to have a method for calculating the likelihood of observing a pair of homologous sequences. For richer models, it is necessary to have a method for calculating the likelihood of observing a pair of homologous sequences. For richer models, it is necessary to have a method for calculating the likelihood of observing a pair of homologous sequences.

To whom correspondence should be addressed.
it possible to infer parameter, either by maximum likelihood
or in a Bayesian fashion. Our approach has the advantage of
using an explicit evolutionary (‘process-based’) model, and
takes into account the equilibrium sequence distribution as a
function of the substitution rate parameters. We demonstrate
the method by estimating substitution rates and confidence
intervals on non-coding human–mouse data. The algorithms
involve some approximations, and we show by experiments on
synthetic data that mutation rates can be faithfully recovered,
using a Bayesian MCMC sampling approach, in the parameter
range corresponding to human–mouse data.

In contrast to most stochastic models used in evolution-
ary biology, the proposed model is naturally irreversible.
Reversible models enjoy technical advantages, for instance,
they have approximately half as many parameters as irrevers-
able models, and have symmetry properties that are helpful
for deriving properties of such models, and in practical
computations. For example, Felsenstein (1981) coined the
Pulley Principle, which states that the likelihood of sequences
evolving according to a reversible substitution model on a
phylogenetic tree is independent of the position of the root,
so that root placement is only possible using an outgroup as
reference. However, there is no a priori reason to assume
reversibility, since many biological processes have a distinct
direction in time, and this is certainly true for evolutionary
processes. The possibility of rooting trees under irreversible
models of substitution was noted before, see e.g. Yang (1994),
but for single nucleotide models the signal seems to be weak,
especially in non-coding DNA (data not shown). The proposed
dinucleotide model incorporates the profoundly directional
CpG effect, making the model strongly irreversible, and we
show that it is possible to infer root positions, even for just
two sequences.

The paper is organized as follows. First, we introduce
the model and discuss some of its properties. We then use
Bayesian MCMC sampling to infer the model parameters.
Next, the method is validated by inferring parameters from
synthetic data. The same procedure is then used on two sets
of 100 kb non-coding human–mouse aligned sequence data
from human chromosomes 21 and 10. The Discussion section
concludes the paper. Finally in the Appendix, we formally
define the proposed model and derive the algorithms for com-
puting the equilibrium distribution, the sequence-to-sequence
likelihood, and the likelihood that two sequences have evolved
from an unknown common ancestor.

THE DINUCLEOTIDE SUBSTITUTION MODEL

We now introduce the ‘dinucleotide model’, a continuous-
time Markov model for nucleotide substitutions. The para-
eters of the model are given by a $16 \times 16$ rate matrix $M$,
whose rows and columns are labelled by the 16 possible nuc-
etide pairs, so that the matrix describes mutation rates from
any nucleotide pair to any other. These rates apply to each of

$$ M : \begin{bmatrix} R_0 : & \begin{bmatrix} 1 & \ldots & 1 & \ldots & 1 \end{bmatrix} \end{bmatrix} \quad R : \begin{bmatrix} M & M \end{bmatrix} $$

Fig. 1. Illustration of the dinucleotide model. Horizontal bars indicate instantaneous (rate) dependencies, grey areas indicate regions of finite-time dependencies due to ‘contagious dependence’. The model is parameterized by a $16 \times 16$ matrix, $M$, specifying mutation rates upon dinucleotides. The matrix $R_i$ has dimension $4^2 \times 4^2$, and corresponds to $M$ acting on dinucleotides $k$ and $k + 1$ only, with no mutation process acting on any other nucleotides. Formally, it is the ‘matrix concatenation sum’ of the null matrix acting on the leftmost $k - 1$ dinucleotides, the matrix $M$, and the null matrix acting on the remaining $L - k - 1$ nucleotides (see Appendix). The full model has rate matrix $R = \sum_{k=1}^{L-1} R_k$, corresponding to the dinucleotide substitution process acting on all $L - 1$ dinucleotides simultaneously.

the $L - 1$ pairs of neighbouring nucleotides in a sequence of
length $L$ simultaneously (Fig. 1). The rate matrix of the full
model, denoted by $R$, specifies rates at which any length-$L$
sequence mutates into any other. This matrix has dimension
$4^L \times 4^L$, but is very sparse; in fact $R_{\sigma,\tau}$, the rate at which
sequence $\sigma$ mutates into $\tau$, vanishes unless $\sigma$ and $\tau$
coincide apart from at most two consecutive nucleotides.

The dinucleotide substitution model introduces dependen-
cies between neighbouring sites, and the stationary sequence
distribution $\pi(\sigma)$ no longer factorizes into a product of single-
nucleotide distributions as in the independent-site model (see
Appendix for an algorithm to compute the stationary distri-
bution). The relation between the parameters of the model (the
coefficients of $M$) and the reversibility of $R$ is more complic-
ted than for independent-nucleotide models, as it involves
this equilibrium sequence distribution. Even for a reversible
$M$ (on length-2 sequences), the total matrix $R$ is in general
irreversible. For example, $M$ may specify detailed balance for
C$G$ ↔ $T$G state transitions if confined to length-2 sequences,
but state transitions of longer sequences that involve mutations
overlapping the $C$ or $G$ residue may disrupt detailed balance
by creating additional $C$G dinucleotides, leading to cycles in
the equilibrium flow graph (Fig. 2).

The matrix $R$ is far too big to use explicitly. It turns out
that it is possible to compute $\exp(R_{\sigma,\tau})$, the probability that
sequence $\sigma$ evolves into $\tau$ in time $t$, without computing the
matrix exponential explicitly, through a dynamic program-
ming recursion that uses the structure of $R$. Exact results
still involve large matrices, and approximations are neces-
sary. Our approximation consists of ignoring all terms related
to multiple substitutions involving four or more consecutive
nucleotides. Such events comprise at least three independent
‘overlapping’ substitutions, so that the leading error term is
cubic in the divergence time and mutation rate. To validate
the approximation in the parameter range of interest, we do
parameter inference on synthetic data.
Fig. 2. Example of irreversibility in the dinucleotide model. Depicted is part of the full Markov chain for sequences of length 4. In this example, rates for the mutation of CG into TG or CA are both 1.0 mutations per observed pair and time unit, while every other neighbour-dependent mononucleotide substitution occurs with a rate of 0.1. The resulting equilibrium probabilities for the length-4 sequences are shown between brackets (see Appendix), and equilibrium flows (in units of $10^{-4}$ transitions per unit of time) are shown alongside the arrows, which point in the direction of net flow. Two rate parameters contribute to each single nucleotide substitution rate, e.g. both CG → TG and GC → GT contribute to the GCGT → GTGT transition, so that the net flow at equilibrium along the edge GCGT → GTGT is 0.00160 × $(1.0 + 0.1) - 0.00475 × (0.1 + 0.1) = 0.00081$. This violation of ‘detailed balance’ implies irreversibility; e.g. the cycle GCGT → GTGT → GGGT → GCGT is more probable to occur than its reversal, giving a definite direction to time.

EVALUATION AND RESULTS

For the substitution model, we used only a subset of the 240 free parameters in the matrix $M$. The symmetry of the substitution process over reverse-complement means that all mononucleotide substitutions can be described by the $4 \times 4 \times 3 = 48$ right-neighbour rates only. General dinucleotide substitutions would require another 80 parameters, but since such substitutions are rare, reliable parameter inference requires much input data, and for this reason we use a single dinucleotide substitution rate parameter, 49 parameters in all.

Synthetic sequence data were produced by simulating the dinucleotide substitution model on a 100 kb sequence. We chose parameters to roughly mimic the parameters expected for human–mouse data, namely, a mononucleotide substitution rate of 0.075 for all substitutions except CG → TG (and CA) which occur with a rate of 2.4. Summing over the implied equilibrium sequence distribution yields a total mononucleotide substitution rate of 0.502 substitutions per site and unit of time. We chose a total dinucleotide substitution rate of 0.020 dinucleotide substitutions per site and unit of time. Since about half as many substitutions have occurred in humans compared with mice since divergence (Mouse Genome Sequencing Consortium, 2002), we chose the root position to be 0.3 time units from the ‘human’ descendant and 0.7 units from the ‘mouse’ sequence.

Neutrally evolving aligned human–mouse sequence data was prepared from BlastZ-aligned data (ftp://genome.ucsc.edu/goldenPath/hg18/). We applied a simple but stringent syneny filter to remove any spurious hits, then removed alignments that overlapped with genes (including introns and regulatory elements), which included repeats (both transposons and tandem repeats), or for which the DUST program (cut-off 16) annotated part of the alignment as a low entropy region. We further removed CpG islands (defined as 250 bp windows containing in excess of 7.5% CpGs, including their 125 bp shoulders; this removed 1.0% of sequence). The remaining data were cut into individual ungapped alignments. Since there is evidence that sequences shorter than ~12 nt cannot always be aligned correctly (data not shown), we trimmed the alignments by removing the leading and trailing 12 nt, and subsequently removed alignments of <10 bases. Finally, we randomly selected a ~100 kb subset of the resulting alignments. This procedure was carried out for human chromosomes 21 (101 142 nt) and 10 (99 563 nt).

RESULTS

Parameter estimation was carried out by Bayesian MCMC sampling running for 600,000 iterations, using flat priors for all parameters. Estimated sample sizes were at least 300–500 for the log-likelihood and typically 100 for the various matrix entries.

The rate estimates from synthetic data are shown in Figure 3a. The estimated total mono- and dinucleotide rates are within 1 SD of their true values. This is also true for >80% of the matrix entries, including the CG → TG rate parameter, suggesting that the estimation method is unbiased. The CG → AG and CG → GG rates come out high, probably due to a combination of crosstalk from the high CG → TG rate and the three-site approximation we use; with a lower CG → TG rate no bias was observed (data not shown). The estimated posterior density for the root position is shown in Figure 4a. The true root position is within 1 SD of the Bayesian estimate of 0.33 ± 0.03.

Rate estimates based on human chromosomes 21 (C21) and 10 (C10) data are shown in Figure 3b and c. The estimates for the two chromosomes are broadly similar. The CG → TG rates are higher than the average mononucleotide rates by a factor 18 (C21; CpG abundance 0.93%) and 17 (C10; CpG abundance 1.06%). The total effective substitution rate for C21, due to mononucleotide and dinucleotide substitutions, is $0.469 ± 0.016 = 0.501$. Of this, $9.4 ± 0.5\%$ is due to the CpG effect, and a further $6.4 ± 0.8\%$ is due to dinucleotide substitutions. For C10, the total rate is 0.487, of which $10.0 ± 0.5\%$ is due to the CpG effect and $6.2 ± 0.8\%$ to dinucleotide substitutions.

Root positions for chromosomes 21 and 10 were estimated at 0.484 ± 0.014 and 0.510 ± 0.016, respectively. Figure 4b plots the posterior densities for both chromosomes.

Figure 5 gives a re-parameterized view of the rate estimates obtained by separating out the neighbour-independent and neighbour-dependent substitution rates. For the synthetic data, the latter are theoretically zero, but since rates are non-negative, they have a non-Gaussian distribution.
A nucleotide substitution model

Fig. 3. Estimated mononucleotide substitution rates (dependent on unchanged right neighbour (top row); left-neighbour dependent rates are fixed by strand reversal symmetry), total mononucleotide rate ($\rho_1$) and total dinucleotide rate ($\rho_2$). Superscripts indicate 1 SD in the last digit(s). (a) Synthetic data; true mononucleotide rates: CG $\rightarrow$ TG, 2.40; all others, 0.075. (b) Chromosomes 21 and (c) 10.

Fig. 4. Posterior density estimates of the root position. (a) The results for synthetic data. The theoretical posterior (with rate matrix fixed to correct values) is shown for comparison (dotted line); the smooth curve is the log-likelihood. The sampled posterior is slightly broadened, due to the co-sampling of rates together with the root position parameter. (b) The results for chromosomes 21 (solid line) and 10 (dotted line).

with non-zero mean. We used this parameterization to test neighbour-dependence, by using synthetic data to estimate cut-off values for the neighbour-independent rates relative to their empirical SD. A cut-off of 2.2 empirical SDs was found to correspond to a 90% confidence level. As expected, the hypothesis of neighbour-independence can be rejected for the CG $\rightarrow$ TG substitution, and indeed for many more.

DISCUSSION

We have introduced a context-dependent substitution model that enables direct estimates of neighbour-dependent and dinucleotide substitution rates. The model is furthermore time-irreversible, which allows root placement in the absence of an outgroup.

We found strong CG $\rightarrow$ TG and CA substitution rates as expected, 17 and 18 times above the average rate for other dinucleotides, in agreement with the previous estimates of a 10–20, fold increase (Sved and Bird, 1990). Our results indicate that the CpG-related substitutions accounts for about ~10% of all substitutions, while an estimate by Subramanian and Kumar (2003) puts the CpG contribution to point substitutions in primate intergenic DNA to ~20%. This 2-fold difference may be partly explained by a different balance of ordinary versus CpG mutations in primates compared with rodents. In concordance with this hypothesis, we find a lower incidence of CpG’s in our human chromosome 21 data-set compared with mouse, although in chromosome 10, the proportions are similar.

The inferred relative contribution of dinucleotide substitutions to the overall per-site substitution rate of ~6% in presumably neutrally evolving human–mouse DNA is in broad agreement to a study by Averof et al. (2000), who reported...
Fig. 5. Testing neighbour dependence of mononucleotide substitutions. The first matrix tabulates the neighbour-independent contribution to the substitution rates (row, original; column, mutant), the other four tabulate rates depending on the (unchanged) right neighbour (indicated at top). For each of these rates, the sample average was compared with the estimated SD to indicate the confidence level at which the zero-rate hypothesis can be rejected (indicated by colours; white corresponds to a 90% level threshold as calibrated on synthetic data). (a) Synthetic data. Only the CG → TG rate is significantly non-zero, as expected. (b) The results for chromosomes 21 and 10. a figure equivalent to 4%. However, Smith et al. (2003) convincingly argued that this estimate could be upwardly biased by rate variation along the genome, an effect we did not include, but is known to be important. A partial filtering for such rate variation resulted in a 2-fold reduction in the dinucleotide rate estimates (data not shown), suggesting that the figure of 6% is an overestimate.

The inferred root position is almost halfway the human and mouse tips. This is surprising since the mouse lineage had attracted about twice as many point mutations as the human lineage since divergence (Mouse Genome Sequencing Consortium, 2002). Since the root inference is based solely on the irreversible signal in the data, one possible explanation is that the mutation processes in mouse and human are not identical, even after scaling, but are a combination of an evolutionarily relatively constant irreversible process and a scaled reversible process reponsible for the majority of observed mutations. This hypothesis can be tested by inferring substitution rates on both lineages independently, using ancestral repeats. The dinucleotide model will hopefully contribute to more precise phylogenetic estimates, by the ability of root inference, and its more accurate modelling of the neutral substitution process. We also intend to use it for a more accurate estimate of the proportion of the human genome under purifying selection, which is currently estimated at 5% (Mouse Genome Sequencing Consortium, 2002). Finally, it may find application in the evolutionary modelling of RNA base stacking, where context dependencies are known to be important.

**ACKNOWLEDGEMENTS**

The authors thank Alexei Drummond, Ian Holmes, Jens-Leedt Jensen, Bjarne Knudsen, István Miklós, Yun Song, Simon Whelan and Ziheng Yang for helpful discussions and valuable suggestions, and two anonymous referees for their careful reading and useful remarks.

**REFERENCES**


APPENDIX

Formal definition of the model
To describe the model more formally, we introduce some notation. Let \( \Omega = \{A, C, G, T\} \) be the alphabet and \( \Omega^2 \) the state space of sequences. The space of probability distributions over \( \Omega^2 \) is denoted by \( \mathcal{D}_L \subset \mathbb{R}^{L^2} \), and a probability distribution \( v \in \mathcal{D}_L \) is a vector assigning a probability to all \( 4^L \) possible sequences in \( \Omega^L \). We label the coordinates of \( \mathcal{D}_L \) by sequences, so that if \( v \in \mathcal{D}_L \) and \( \sigma \in \Omega^L \), \( v_\sigma \) is the probability of observing the sequence \( \sigma \). Similarly, for a matrix \( A \), a matrix coefficient is written \( A_{\sigma,\tau} \), and is interpreted as the rate at which sequence \( \sigma \) mutates into \( \tau \) for rate matrices, or the probability that sequence \( \sigma \) mutates into \( \tau \) for probability matrices. We write \( \sigma \tau \) for the concatenation of \( \sigma \) and \( \tau \), and we write \( \sigma[i:j] \) for the subsequence \( \sigma[i] \cdots \sigma[j] \). For rate matrices \( A, B \) acting on \( \mathcal{D}_k \) and \( \mathcal{D}_l \), respectively, we denote by \( A \oplus B \) (the matrix concatenation sum of \( A \) and \( B \)) the matrix acting on \( \mathcal{D}_{k+l} \) that has \( A \) acting on the leftmost \( k \) residues of the sequence, so that it neither depends on nor changes the rightmost \( l \) residues, while \( B \) independently and simultaneously acts on the rightmost \( l \) residues. In particular, this operation is not commutative: \( A \oplus B \neq B \oplus A \), however it is associative, \( A \oplus (B \oplus C) = (A \oplus B) \oplus C \). Formally, \( (A \oplus B)_{\sigma \tau q} = A_{\rho \sigma} B_{\tau q} + B_{\rho \tau} A_{\sigma q} \), where \( B_{\sigma \tau} = 1 \) if \( \sigma = \tau \) and 0 otherwise. For example, \((A\oplus B)_{psql} = 0\) for \( p \neq q \) and \( s \neq t \), since the rate for two independent mutations to occur simultaneously vanishes. (Note that this matrix concatenation sum is distinct from the direct sum of matrices, for which the same symbol \( \oplus \) is commonly used.) Finally, let \( O_k \) be the null matrix on \( \mathcal{D}_k \), then the rate matrix for the dinucleotide model on a sequence of length \( L \) is

\[
\mathcal{R} := \mathcal{R}_L = \sum_{k=1}^{L-1} O_{L-1} \oplus M \oplus O_{L-k-1}.
\]

Stationary sequence distribution
The dinucleotide substitution model introduces dependencies between neighbouring sites, and the stationary sequence distribution \( \pi(\sigma) \) no longer factorizes into a product of single-nucleotide distributions as in the independent-site model. For a certain class of reversible dinucleotide substitution models, the stationary distribution is of Gibbs form (Pedersen and Jensen, 2001). It can be shown that this implies a (first order) Markov structure for the stationary distribution, i.e.

\[
p(\sigma_i = \alpha \mid \sigma_1 \sigma_2 \cdots \sigma_{i-1}) = p(\sigma_i = \alpha \mid \sigma_{i-1}).
\]

In general, non-reversible case, numerical experiments seem to indicate that this Markov property breaks down, even though the rate matrix involves only pairwise interactions. It is unclear whether this is a result of the irreversibility of the process, or whether reversibility and having a Markovian stationary distribution are orthogonal features. At any rate, there seems to be no simple expression for the stationary distribution, and we have to resort to a numerical approximation.

The matrix \( \mathcal{R}_K \) can be built explicitly for small \( K \), and we can find its stationary distribution numerically by solving \( \pi \mathcal{R}_K = 0 \). However, this will not properly approximate the marginal distribution \( p(\sigma_1 \cdots \sigma_{i+K-1}) \) for a length-\( K \) subsequence in a longer sequence of length \( L \), because no substitutions overlapping the edges are taken into account. Such edge effects can be taken into account as follows. First, we note that although (2) is not satisfied exactly, a higher-order Markov property does hold approximately,

\[
p(\sigma_i = \alpha \mid \sigma_1 \sigma_2 \cdots \sigma_{i-1}) \approx p(\sigma_i = \alpha \mid \sigma_{i-n} \cdots \sigma_{i-1}),
\]

and the approximation converges exponentially in \( n \). If we know the exact marginal distribution \( \pi \) of length-\( K \) subsequences, we can use (3) to find the approximate conditional distribution of \( \sigma_{K+1} \).

\[
p(\sigma_{K+1} \mid \sigma_1 \cdots \sigma_K) \approx p(\sigma_{K+1} \mid \sigma_2 \cdots \sigma_K) \frac{\pi(\sigma_2 \cdots \sigma_K \sigma_{K+1})}{\sum_{\sigma \in \Omega} \pi(\sigma_2 \cdots \sigma_K \sigma)},
\]

This approximation is known as the ‘\( K \)-cluster approximation’ in the physics literature (Arndt et al., 2003; ben Avraham and Köhler, 1992). We can now include edge effects by having the rate matrix \( M \) act on \( \sigma_K \sigma_{K+1} \) by supposing that \( \sigma_{K+1} \) is distributed according to (4), and similarly for the left-hand edge. Formally, we add to \( \mathcal{R}_K \) the rate matrix \( \mathcal{R}_K' \), describing
the substitutions at the edges:

\[(\mathcal{R}'_K)_{\sigma,\tau} = \left[ \sum_{\alpha,\beta} \pi(\alpha \sigma_1 \cdots \sigma_{K-1}) M_{\alpha \sigma_{K-1}, \beta} + \sum_{\alpha} \pi(\sigma_2 \cdots \sigma_K) M_{\alpha \sigma_{K-1}, \tau} \right] \delta_{\sigma_1[2:K], \tau[2:K]} \times \delta_{\sigma[1:K-1], \tau[1:K-1]} \right], \tag{5}

where \(\sigma\) and \(\tau\) are sequences in \(\Omega^K\). From an initial guess for \(\pi\), we compute \(\mathcal{R}'_K\), and then solve \(\pi(\mathcal{R}_K + \mathcal{R}'_K) = 0\) for \(\pi\) to get a better approximation. This procedure is repeated until convergence, which is rapid. The only approximation is made in (4), and since the correlation between nucleotides decreases exponentially fast with their separation, this approximation can be good even for moderate values of \(K\). In this paper, we use \(K = 3\).

A recursion for sequence-to-sequence probabilities

Let \(v(t)\) be the probability distribution vector at time \(t\), so that \(v(t)\) is the probability of observing sequence \(\sigma\) at time \(t\). Since the rate at which sequence \(\sigma\) mutates into sequence \(\tau\) is \(\mathcal{R}_{\sigma,\tau}\), the time evolution of \(v\) is given by \(dv(t)/dt = v(t)\mathcal{R}\). The solution to this equation is \(v(t) = v(0) \exp(\mathcal{R}t)\), and the probability of sequence \(\sigma\) evolving into \(\tau\) in time \(t\) is \(\exp(\mathcal{R}t)_{\sigma,\tau}\). However, the matrix \(\mathcal{R}\) is of dimension \(4^L\), too big for explicit computations. Write \(R_i := O_{-1} + M + O_{L-i-1}\), and recall that \(\mathcal{R} = \sum_{i=1}^{L-1} R_i\). We may expand the matrix exponential in a Taylor series,

\[\exp(\mathcal{R}t) = 1 + \left( \sum_{i=1}^{L-1} R_i \right) t + \frac{1}{2!} \left( \sum_{i=1}^{L-1} R_i \right)^2 t^2 + \cdots. \tag{6}\]

Many of the terms in the expression \((\sum R_i)^n\) commute; indeed, \(R_i R_j = R_j R_i\) unless \(|i - j| = 1\). We say that a factor \(R_i\) is overlapping if it cannot be written as the product of two commuting factors. For instance, \(R_1 R_2 R_3\) is overlapping, but \(R_1 R_2 R_3\) is not, since by swapping the middle two factors (which commute), we get \((R_1 R_2)(R_3 R_5)\), a product of two commuting factors. In this way, a term can be written uniquely as a product of commuting factors, which themselves are overlapping. We define the length of an overlapping factor to be the number of sites it affects, e.g. the length of \(R_1 R_2 R_3\) is 3 if it affects sites \(1\) through \(3\).

Now if a pair of neighbouring sites has never experienced a substitution involving both nucleotides simultaneously, the evolutionary histories of the left and right sequence parts become independent, and the likelihood factorizes into a product. If we expand the full likelihood in terms of the first position (counted from the right) where such a ‘break’ in the dependence structure occurred, we obtain a dynamic programming recursion.

Mathematically, we factorize the terms of (6) into commuting factors. Consider all terms that contain in their factorization an overlapping factor \(F = R_1 \cdots R_n\) of length \(k\) that includes a factor \(R_{i-1}\). The sum of these terms can be written as \(GF\), and this product commutes by construction; \(F\) only contains terms \(R_i\) with \(i > L - k\), whereas \(G\) only contains \(i < L - k\) terms. In fact, we have

\[G = I_{L-k} - \frac{t^n}{n!} \left( \sum_{i=1}^{L-k-1} R_i \right) \left( \frac{n+1}{(n+1)!} \right) ^{t^n + \cdots}

= \frac{t^n}{n!} \left[ I_{L-k} - \left( \sum_{i=1}^{L-k-1} R_i \right) \frac{t^n}{n!} \right] ^{2 \frac{t^2}{2!} + \cdots}

= \frac{t^n}{n!} \exp \left[ \mathcal{R}_{L-k} \otimes \Omega \right] t^n

= \frac{t^n}{n!} \exp(\mathcal{R}_{L-k}) \otimes I_k, \tag{7}\]

where the binomial coefficients \(\binom{n+k}{k}\) count the number of ways that \(k\) factors \(R_i\) can be interleaved with the \(n\) factors comprising \(F\) in the product \((\sum R_i)^{n+k}\). Here, we introduced the matrix concatenation product, \(\otimes\), which is defined by \((A \otimes B)(x,y) = A x B\), and the symbol \(I_k\) denotes the identity matrix on \(D_k\). Recall that \(\mathcal{R}_k\) is the rate matrix acting on \(D_k\) as defined in (1). If we denote by \(A_k\) the sum of all overlapping factors \(F\) of length \(k\), including a factor \(t^n/n!\) each, then from (7) it follows that

\[\exp(\mathcal{R}_{n,t}) = e^{R_{n-1,t}} \otimes A_1 + e^{R_{n-1,t}} \otimes A_2 + \cdots + A_n. \tag{8}\]

(Here, we included the identity matrix \(I_1\) into \(A_1\).) Now, let \(P_n\) be the probability that the length-\(n\) prefix of \(\sigma\) evolves into the same prefix of \(\tau\). More formally, \(P_n = [\exp(\mathcal{R}_{n,t})]_{\sigma[1:n],\tau[1:n]}\), where we introduced the notation \(\sigma[i,j] = \sigma_i \sigma_{i+1} \cdots \sigma_j\). Then, we can turn (8) into the following dynamic programming recursion:

\[P_n = (A_1)_{\sigma[1:n-1],\tau[1:n-1]} + (A_2)_{\sigma[1:n-2],\tau[1:n-2]} + (A_3)_{\sigma[1:n-3],\tau[1:n-3]} + \cdots, \tag{9}\]

with the initialization \(P_0 = 1\). To compute the \(A_k\), we iteratively solve for \(A_1, A_2, \ldots\) in (8). For \(n = 1\), the equation reads \(\exp(\mathcal{R}_{1,t}) = A_1\), and there is nothing to solve. Note that \(\mathcal{R}_1 = 0\) by definition (1), so that \(A_1 = I_1\). The other factors are found recursively:

\[A_2 = e^{R_{2,t}} - e^{R_{2,t}} \otimes A_1, \tag{10}\]

\[A_3 = e^{R_{3,t}} - e^{R_{3,t}} \otimes A_1 - e^{R_{3,t}} \otimes A_2, \tag{11}\]

\[A_4 = e^{R_{4,t}} - e^{R_{4,t}} \otimes A_1 - e^{R_{4,t}} \otimes A_2 - e^{R_{4,t}} \otimes A_3. \tag{12}\]
If these formulas are expanded in terms of the $A_k$, we get
\begin{align}
A_2 &= e^{R_2} - A_1 \otimes A_1, \\
A_3 &= e^{R_3} - A_2 \otimes A_1 - A_1 \otimes A_2 - A_1 \otimes A_1 \otimes A_1, \\
A_4 &= e^{R_4} - A_3 \otimes A_1 - A_2 \otimes A_2 - A_1 \otimes A_3 \\
&- A_2 \otimes A_1 \otimes A_1 - A_1 \otimes A_2 \otimes A_1 \\
&- A_1 \otimes A_1 \otimes A_2 - A_1 \otimes A_1 \otimes A_1 \otimes A_1. 
\end{align}
(13)
(14)
(15)

Collected on the right-hand sides are all possible ways in which a matrix on $D_k$ can be built from a matrix concatenation product of matrices $A_i$, $i < k$. By definition, the terms occurring in such products are not overlapping. Since the $A_i$ contain all overlapping terms of length $i$ in the expansion of $\exp(R_i)$, the terms in $A_k$ are those in $\exp(R_k)$ except terms that factorize, i.e. all overlapping terms of length-$k$.

The recursion (9) is exact, but in practise only a few terms can be included, since the dimension of the matrices $A_i$ grows exponentially with $i$. Fortunately, the matrix entries tend to 0 exponentially fast, and a good approximation can be obtained with a few terms. In the implementation, we used the Padé algorithm to compute the matrix exponentials of the non-symmetric matrices (see Moler and van Loan, 2003).

**Evolution from a common ancestor**

The algorithm developed above computes the likelihood that one sequence evolves into another. Most often however, we require this independence to hold on both branches simultaneously. Let $p^n_{\sigma, \tau}$ be the likelihood of the descendant sequence prefixes $\sigma[1, n]$ and $\tau[1, n]$ to have evolved from a common ancestral sequence prefix $\rho[1, n]$ in time $t_1$, $t_2$, respectively, where the unobserved ancestral sequence is distributed according to the equilibrium distribution, conditional on the last two nucleotides $\rho_{n-1} \rho_n$ being $\beta \alpha$. Analogous to (9) we then have the following dynamic programming recursion:

\begin{align}
P^\beta_\alpha &_{n} = \sum_\gamma P^\gamma_\beta P(\gamma | \beta \alpha) B^\alpha_{\sigma, \tau} \\
&+ \sum_\gamma P^\gamma_\beta P(\delta | \gamma \beta) B^\beta_\gamma \sigma_{[n-2, n], \tau_{[n-2, n]}} \\
&+ \sum_\gamma P^\gamma_\beta P(\epsilon | \delta \gamma) B^\epsilon_\delta \beta_{\sigma_{[n-2, n], \tau_{[n-2, n]}}}, 
\end{align}
(16)

Here, $p(\gamma | \beta \alpha) = \pi(\gamma \beta \alpha) / \sum_\delta \pi(\delta \beta \alpha)$ is the probability of observing $\gamma$ conditional on its right neighbours $\beta \alpha$. This recursion can be made more efficient, removing the double and triple summations, by expressing the stationary distribution in terms of a nucleotide pair further up along the sequence:

\begin{align}
P^\beta_\alpha &_{n, 0} = P^\beta_\alpha_{n-1} B^\beta_\alpha_{\sigma_{[n-1, n]], \tau_{[n-1, n]}} \\
&+ \sum_\gamma P^\gamma_\beta P(\gamma | \beta \alpha) B^\gamma_\beta \sigma_{[n-2, n], \tau_{[n-2, n]}}, \\
P^\beta_\alpha &_{n, k+1} = \sum_\gamma P^\gamma_\beta p(\gamma | \beta \alpha) \quad (k = 0, 1).
\end{align}
(17)
(18)

The $B$-factors represent the probabilities of the events that yield the required dependencies on the two branches, and can be computed by a procedure similar to that used for the sequence-to-sequence case:

\begin{align}
B^\sigma_{\sigma_{[n, n]}, t_{1}} &= (e^{R_{t_{1}}} \sigma_{[n, n]}) B^\sigma_{\sigma_{[n, n]}, t_{1}}, \\
B^\beta_{\sigma_{[n, n]}, t_{2}} &= (e^{R_{t_{2}}} \beta_{\sigma_{[n, n]}, t_{2}}) B^\beta_{\sigma_{[n, n]}, t_{2}} - B^\beta_{\sigma_{[n, n]}, t_{2}} B^\beta_{\sigma_{[n, n]}, t_{2}}, \\
B^{\gamma \beta}_{\sigma_{[n, n]}, t_{1}, t_{2}} &= (e^{R_{t_{1}}} \gamma \beta_{\sigma_{[n, n]}, t_{1}, t_{2}}) B^{\gamma \beta}_{\sigma_{[n, n]}, t_{1}, t_{2}}, \\
&- B^{\gamma \beta}_{\sigma_{[n, n]}, t_{1}, t_{2}} B^{\beta \gamma}_{\sigma_{[n, n]}, t_{1}, t_{2}} B^{\beta \gamma}_{\sigma_{[n, n]}, t_{1}, t_{2}}, \\
&- B^{\gamma \beta}_{\sigma_{[n, n]}, t_{1}, t_{2}} B^{\beta \gamma}_{\sigma_{[n, n]}, t_{1}, t_{2}},
\end{align}
(19)

To initialize the recursion, we deviate slightly from the model and assume that the length-$L$ sequence is embedded in an infinitely long sequence. This ensures that the nucleotides at the edges are subjected to the same mutation rates as nucleotides at other positions. This idea is implemented by setting $P^\beta_\alpha = 1$ for $i < 1$, and summing over the unobserved nucleotides $\sigma_i$ and $\tau_i$ with $i < 1$ in (16). At termination, the recursion (16) is executed for two additional steps, up to $n = L + 2$, similarly summing over the unobserved nucleotides $\sigma_i$ and $\tau_i$ with $i > L$. Finally, the likelihood is $P_{\sigma, \tau} = \sum_{\sigma, \tau} \pi(\beta \alpha) P^\beta_\alpha_{L+2}.$