2005

Evolutionary models for insertions and deletions in a probabilistic modeling framework

Elena Rivas
Washington University School of Medicine in St. Louis

BMC Bioinformatics

Research article

Evolutionary models for insertions and deletions in a probabilistic modeling framework

Elena Rivas*

Address: Department of Genetics, Washington University School of Medicine, 4444 Forest Park Blvd., Saint Louis, Missouri 63108 USA

Email: Elena Rivas* - elena@genetics.wustl.edu

* Corresponding author

Published: 21 March 2005

BMC Bioinformatics 2005, **6**:63 doi:10.1186/1471-2105-6-63

Received: 16 December 2004

Accepted: 21 March 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/63

Abstract

Background: Probabilistic models for sequence comparison (such as hidden Markov models and pair hidden Markov models for proteins and mRNAs, or their context-free grammar counterparts for structural RNAs) often assume a fixed degree of divergence. Ideally we would like these models to be conditional on evolutionary divergence time.

Probabilistic models of substitution events are well established, but there has not been a completely satisfactory theoretical framework for modeling insertion and deletion events.

Results: I have developed a method for extending standard Markov substitution models to include gap characters, and another method for the evolution of state transition probabilities in a probabilistic model. These methods use instantaneous rate matrices in a way that is more general than those used for substitution processes, and are sufficient to provide time-dependent models for standard linear and affine gap penalties, respectively.

Given a probabilistic model, we can make all of its emission probabilities (including gap characters) and all its transition probabilities conditional on a chosen divergence time. To do this, we only need to know the parameters of the model at one particular divergence time instance, as well as the parameters of the model at the two extremes of zero and infinite divergence.

I have implemented these methods in a new generation of the RNA genefinder QRNA (eQRNA).

Conclusion: These methods can be applied to incorporate evolutionary models of insertions and deletions into any hidden Markov model or stochastic context-free grammar, in a pair or profile form, for sequence modeling.

Background

Probabilistic models are widely used for sequence analysis [1]. Hidden Markov models (HMMs) are a very large class of probabilistic models used for many problems in biological sequence analysis such as sequence homology searches [2-4], sequence alignment [5], or protein genefinding [6-8]. Stochastic context-free grammars (SCFGs) are another class of probabilistic models used for structural RNAs for problems such as RNA homology searches [9-13], RNA structure prediction [14,15], and RNA genefinding [16].

Sequence similarity methods based on HMMs or SCFGs can take the form of profile or pair models and are very

important for comparative genomics. These probabilistic methods for sequence comparison assume a certain degree of sequence divergence. For instance, in profile models (either profile HMMs [2-4] or profile SCFGs [12,13]) a sequence is compared to a consensus model. Profile models must allow for the occurrence of insertions and deletions with respect to the consensus, and they do so by using state transition probabilities that assign some position-dependent penalties for modifying the consensus with insertions or deletions. Similarly, in pair probabilistic models [8,16] two related sequences are compared (aligned and/or scored). Pairwise alignments need to allow for substitution, insertion and deletion events between the two related sequences. Substitutions are taken care of by residue emission probabilities, while insertion and deletion events are generally taken care of by state transition probabilities as in the case of profile HMMs.

In the BLAST programs [17], the score of a pairwise alignment is determined using substitution matrices which measure the degree of similarity between two aligned residues. Similarly, in pair probabilistic models, residue emission probabilities are based on substitution matrices. The evolution of substitution matrices has been studied at large for many different kinds of processes: nucleotides, amino acids, codons, or RNA basepairs [18-23]. The evolution of emission probabilities using substitution matrices is easily integrated into probabilistic models both for HMMs [24-29] and for SCFGs [14].

In probabilistic models, insertion and deletion events (indels) are sometimes described by treating indels as an additional residue (gap characters) in a substitution matrix. More often they are described using additional hidden states, where transition probabilities into those states represent the cost of gap initiation and transitions within those states represent the cost of gap extension. If the cost of gap initiation and gap extension are identical, it is referred to as a linear gap cost model. Hidden states allow arbitrary costs for gap initiation and gap extension, which is traditionally referred to as an affine gap cost model. Treating gaps as an extra character in a substitution matrix is equivalent to assuming a linear gap cost model. The parameters that modulate those processes should be allowed to change as the divergence time for the sequences being compared is varied. It has been difficult to combine probabilistic models such as profile and pair HMMs or SCFGs with evolutionary models for insertion and deletions [30-33]. Methods to evolve transition probabilities are not as well developed as those describing substitution matrices, but significant effort is currently aimed at this problem [34-41]. Models incorporating the evolution of insertions and deletions in the context of probabilistic models such as profile HMMs or pair models are a very important goal in order to make those probabilistic models more realistic.

I encountered this problem in working on QRNA, a computational program to identify noncoding RNA genes de novo. QRNA uses probabilistic comparative methods to analyze the pattern of mutation present in a pairwise alignment in order to decide whether the compared nucleic acid sequences are more likely to be protein-coding, structural RNA encoding, or neither. Originally QRNA was parameterized at a fixed divergence time. Motivated by the goal of making QRNA a time dependent parametric family of models, I investigated the possibility of evolving the transition and emission probabilities associated with a given probabilistic model. Since I already had the model parameterized for a given time, I aimed to use that model as a generating point of the whole time-parameterized family of models.

Because QRNA includes both linear and affine gap models in different places, in this paper I propose algorithms to describe the evolution of indels as a $(N + 1)$-th character in a substitution matrix, and algorithms to describe the evolution of the transition probabilities associated with a probabilistic model.

The purpose of this paper is to describe the general theoretical framework behind these methods. A detailed description of the particular implementation of these algorithms in QRNA and a discussion of the results obtained with "evolutionary QRNA" (eQRNA) will appear in a complementary publication.

Results

Evolutionary models for emission probabilities

The evolution of emission probabilities without gaps

In order to introduce notation, I will start with a brief review of the current methods for calculating joint probabilities conditional on time, $P(i, j|t)$, where i, j are two residues (for instance, nucleotides, amino acids, RNA basepairs, or codons). $P(i, j|t)$ gives us the probability that residues i and j are observed at a homologous site in a pairwise alignment after a divergence time t. Pairwise sequence comparison methods score aligned residue pairs with these joint probabilities either explicitly or implicitly [17]. In explicit generative pair probabilistic models, like the pair-HMMs and pair-SCFG in QRNA, the $P(i, j|t)$ terms are referred to as pair emission probabilities.

The evolution of joint probabilities is usually obtained by modeling the corresponding conditional probabilities $P(j|i, t)$as a substitution process in which residue i has been substituted by residue j over time t. Probabilistic models for nucleotide substitutions [18,19,42-44] assume

that nucleotide substitution follows a model of evolution that depends on an instantaneous rate matrix,

$$Q_t = e^{tR}, \quad (1)$$

where t is the divergence time, R is the instantaneous rate matrix, and Q_t is the substitution matrix of conditional probabilities, that is $Q_t(ij) \equiv P(j|i, t)$. This is a reasonable model used, for instance, to describe nucleotide substitutions in the Jukes-Cantor [42] or Kimura [43] models, or the more general REV model [44]; this is also the evolutionary model used for amino acid substitutions [18,19,21,45,46], codon to codon substitutions [20,47], and RNA basepair to basepair substitutions [14,22,23].

Throughout this paper, I will use the words "divergence time", "divergence", or "time" equivalently to describe the amount of dissimilarity between biological sequences measured as the number of mutations and gaps introduced in the alignment of the sequences. I will never refer to "time" as representing an actual number of years of divergence, since this number cannot be determined intrinsically from sequence data.

Thus, given a rate matrix R, Q_t (and therefore the desired joint emission probabilities) can be inferred for any desired time using the Taylor expansion for the matrix exponential,

$$Q_t = \sum_{n=0}^{n=\infty} \frac{(tR)^n}{n!} \quad (2)$$
$$= I + tR + \frac{t^2}{2!}RR + \frac{t^3}{3!}RRR + \ldots$$

This Taylor series converges in all cases.

There are several ways in which the rate matrix R can be determined. One approach is to use analytically inferred rate matrices that depend on a small number of external parameters [42-44,48]. For instance, the HKY model for nucleotide substitutions [48] depends on six parameters: the four stationary nucleotide frequencies, a rate of transitions, and a rate of transversions, which have to be provided externally. Another type of approach uses maximum likelihood methods [21,49,50] in order to estimate a rate matrix numerically from a training set of sequence alignments.

A third approach arises naturally in cases where suitable joint probabilities have already been estimated for a pair model, and we wish to make that model conditional on evolutionary divergence time. This approach starts from the assumption that our point estimate represents sequences at a particular arbitrary divergence time t_*. For example, a similar assumption was taken to construct the BLOSUM matrices [51], which were obtained as joint probabilities at discrete point estimates from clusters of aligned sequences.

In this third approach the parameters at the generating time t_* will be used to construct a rate matrix for the process. This approach is motivated by the kind of situation in which we find ourselves with probabilistic methods based on homology such as QRNA: a model has been trained in one kind of data, and the resulting probabilities represent some effective but *fixed* divergence time, and we wish to extend that model to a time-dependent parameterization.

For residue substitution processes, the rate matrix R and Q_*, defined as the substitution matrix at the generating time t_* [$Q_* \equiv Q_{t_*}$], convey exactly the same information. More explicitly, assuming the evolutionary model given in equation (1) we can calculate the rate matrix of the process as a function of Q_* as

$$R = \frac{1}{t_*}\log(Q_*), \quad 0 < t_* < \infty. \quad (3)$$

Kishino *et al* [52] introduced the idea of calculating the rate matrix starting from a given substitution matrix using equation (3) and an eigenvalue decomposition of Q_*. It is worth noting that the matrix equation for the rate R can be expressed as a Taylor expansion of the form

$$R = \frac{1}{t_*}\sum_{n=1}^{n=\infty} \frac{(-1)^{n+1}}{n}(Q_* - I)^n \quad (4)$$
$$= \frac{1}{t_*}\left\{(Q_* - I) - \frac{1}{2}(Q_* - I)^2 + \frac{1}{3}(Q_* - I)^3 - \ldots\right\},$$

which allows for a direct numerical calculation of the rate matrix. The convergence of this series requires only that for every (real or complex) eigenvalue λ of matrix Q_*, then $|\lambda - 1| < 1$. In addition, for any valid substitution matrix the eigenvalues have to be real and $|\lambda| \leq 1$ (see Appendix A). Under these two conditions, the above Taylor series converges so long as the eigenvalues of Q_* are positive. Therefore the three properties required of Q_* in order to be able to obtain a rate matrix using the Taylor expansion in equation (4) are that its eigenvalues are all smaller than one (but one that is strictly one), real, and positive. Complex or negative eigenvalues would correspond to oscillatory behaviors, which do not seem to reflect the biology. All the substitution processes I have tested so far for nucleotides, amino acids, and RNA basepairs correspond to real and positive eigenvalues for which the above method is applicable.

It is relevant to compare instantaneous rate matrix approaches to the approach used in the PAM amino-acid

$$R = \begin{pmatrix}
A & C & D & E & F & G & H & I & K & L & M & N & P & Q & R & S & T & V & W & Y \\
- & .029 & .019 & .056 & .022 & .106 & .011 & .032 & .049 & .088 & .020 & .005 & .045 & .037 & .055 & .238 & .089 & .138 & .003 & .016 \\
.142 & - & .021 & .001 & .024 & .020 & .008 & .056 & .017 & .089 & .028 & .013 & .016 & .012 & .018 & .077 & .058 & .060 & .008 & .019 \\
.026 & .006 & - & .233 & .014 & .052 & .015 & .024 & .034 & .006 & .005 & .100 & .038 & .048 & .013 & .102 & .045 & .014 & .002 & .007 \\
.061 & .000 & .189 & - & .011 & .023 & .030 & .009 & .136 & .018 & .006 & .049 & .035 & .181 & .061 & .096 & .040 & .040 & .003 & .013 \\
.045 & .010 & .022 & .021 & - & .027 & .019 & .111 & .017 & .150 & .051 & .013 & .008 & .006 & .017 & .049 & .033 & .039 & .027 & .209 \\
.113 & .004 & .042 & .023 & .014 & - & .010 & .005 & .024 & .011 & .007 & .061 & .022 & .017 & .025 & .102 & .020 & .020 & .007 & .009 \\
.037 & .006 & .039 & .093 & .031 & .032 & - & .018 & .039 & .022 & .014 & .114 & .025 & .063 & .084 & .060 & .023 & .014 & .005 & .125 \\
.044 & .015 & .024 & .012 & .074 & .006 & .007 & - & .002 & .465 & .054 & .016 & .011 & **-.002** & .017 & .050 & .032 & .692 & .002 & .023 \\
.059 & .004 & .030 & .150 & .010 & .027 & .014 & .002 & - & .049 & .025 & .059 & .040 & .100 & .222 & .113 & .041 & .041 & .003 & .015 \\
.072 & .015 & .004 & .014 & .060 & .008 & .005 & .282 & .033 & - & .138 & .012 & .013 & .021 & .031 & .030 & .053 & .113 & .006 & .025 \\
.069 & .019 & .012 & .020 & .084 & .022 & .014 & .135 & .070 & .566 & - & .026 & .033 & .127 & .070 & .101 & .050 & .239 & .015 & .026 \\
.008 & .004 & .115 & .069 & .010 & .088 & .053 & .018 & .076 & .022 & .012 & - & .018 & .055 & .075 & .191 & .082 & .012 & .001 & .014 \\
.065 & .005 & .041 & .046 & .005 & .031 & .011 & .012 & .048 & .024 & .014 & .017 & - & .034 & .022 & .057 & .047 & .037 & .002 & .009 \\
.068 & .005 & .065 & .307 & .006 & .030 & .035 & **-.002** & .153 & .048 & .070 & .066 & .043 & - & .153 & .124 & .046 & .040 & .008 & .037 \\
.085 & .005 & .015 & .086 & .013 & .036 & .038 & .020 & .282 & .058 & .032 & .075 & .023 & .127 & - & .050 & .059 & .003 & .004 & .017 \\
.233 & .015 & .074 & .087 & .024 & .094 & .018 & .036 & .092 & .036 & .030 & .121 & .039 & .066 & .032 & - & .164 & .005 & .003 & .014 \\
.113 & .015 & .043 & .046 & .021 & .024 & .009 & .030 & .043 & .081 & .019 & .067 & .041 & .032 & .049 & .212 & - & .126 & .007 & .014 \\
.161 & .014 & .012 & .043 & .022 & .022 & .005 & .595 & .040 & .161 & .083 & .009 & .029 & .025 & .002 & .006 & .116 & - & .003 & .031 \\
.020 & .009 & .008 & .018 & .077 & .037 & .008 & .009 & .017 & .043 & .026 & .004 & .008 & .024 & .016 & .018 & .033 & .014 & - & .094 \\
.037 & .009 & .012 & .029 & .241 & .019 & .088 & .040 & .030 & .071 & .018 & .021 & .015 & .047 & .026 & .034 & .027 & .063 & .037 & -
\end{pmatrix}$$

Figure 1
Rate matrix generated from BLOSUM62. Rate matrix obtained from the amino-acid substitution matrix BLOSUM62, rescaled to have an average number of one substitution per amino acid. Notice in bold the two off diagonal negative entries.

substitution matrices [53]. The PAM matrices were not generated by calculating a rate matrix, but by estimating from a collection of highly similar sequences the substitution matrix for the time of one substitution per site $Q_*^{\mathrm{PAM}} \equiv Q_{t=0.01}$, and then calculating Q_t at any other (integer t) time by multiplication. This is a discrete approximation that converges to the same answer given by the rate method for very small time units. PAM matrices have been criticized for not being able to capture the substitutions that are observed for more dissimilar sequences. BLOSUM matrices empirically outperform PAM in sequence homology searches, presumably because sequences at larger divergence times were used to calculate the BLOSUM matrices. However, the BLOSUM method is not a time dependent continuous model but a very coarse-grained discretization. There are ways of combining the best of both approaches (more divergent sequence for training and a continuous-time model) to generate rate matrices, for instance by using the resolvent method [54], or using maximum likelihood methods as in the WAG matrices [21]. However, it is also possible to take a discrete BLOSUM matrix, for instance BLOSUM62, and convert it to an underlying rate matrix. The BLOSUM62-generated rate matrix obtained using equation (4) is shown in Figure 1.

A rate matrix can also be derived from the PAM data Q_*^{PAM} by various methods. One exact method is to do an eigenvalue decomposition as presented in [52]. Recently, other methods have been proposed to calculate a rate matrix from the Dayhoff data [55]. These methods still assume that $R \approx (Q_*^{\mathrm{PAM}} - I)$ which corresponds to taking only the first term in the Taylor series for the logarithm in equation (4). This assumption is good only for very closely related sequences. Using the Taylor series allows one to estimate, using the same input data and avoiding the calculation of eigenvalues, the rate matrix to any desired level of precision, independent of the degree of similarity in the training set.

Notice that the rate matrix obtained using BLOSUM62 (Figure 1) has two off-diagonal negative entries (and if we use more divergent BLOSUM matrices we have more negative off-diagonals). Off-diagonal entries of the rate matrix have to be positive so that $I + \delta t R$ can be interpreted as a substitution matrix for very small times δt. This problem is not unique to sequence data. The construction of rate matrices for a Markov process from empirical data using a generating time is also used in mathematical modeling of financial processes such as credit risk modeling [56,57]. In the world of mathematical finances the problem is referred to as the regularization problem. I will use one of following regularization algorithms presented in [57]. The QOG algorithm (quasi-optimization of the generator) regularizes the rate matrix. The QOM algorithm (quasi-optimization of the root matrix) leaves the rate matrix unchanged and regularizes the conditional matrix at a given time if any negative probability appears. Using

$$R = \begin{pmatrix}
A & C & D & E & F & G & H & I & K & L & M & N & P & Q & R & S & T & V & W & Y \\
- & .029 & .019 & .056 & .022 & .106 & .011 & .032 & .049 & .088 & .020 & .005 & .045 & .037 & .055 & .238 & .089 & .138 & .003 & .016 \\
.142 & - & .021 & .001 & .024 & .020 & .008 & .056 & .017 & .089 & .028 & .013 & .016 & .012 & .018 & .077 & .058 & .060 & .008 & .019 \\
.026 & .006 & - & .233 & .014 & .052 & .015 & .024 & .034 & .006 & .005 & .100 & .038 & .048 & .013 & .102 & .045 & .014 & .002 & .007 \\
.061 & .000 & .189 & - & .011 & .023 & .030 & .009 & .136 & .018 & .006 & .049 & .035 & .181 & .061 & .096 & .040 & .040 & .003 & .013 \\
.045 & .010 & .022 & .021 & - & .027 & .019 & .111 & .017 & .150 & .051 & .013 & .008 & .006 & .017 & .049 & .033 & .039 & .027 & .209 \\
.113 & .004 & .042 & .023 & .014 & - & .010 & .005 & .024 & .011 & .007 & .061 & .022 & .017 & .025 & .102 & .020 & .020 & .007 & .009 \\
.037 & .006 & .039 & .093 & .031 & .032 & - & .018 & .039 & .022 & .014 & .114 & .025 & .063 & .084 & .060 & .023 & .014 & .005 & .125 \\
.044 & .015 & .024 & .011 & .074 & .006 & .007 & - & .002 & .465 & .054 & .015 & .011 & **.000** & .017 & .049 & .032 & .692 & .002 & .023 \\
.059 & .004 & .030 & .150 & .010 & .027 & .014 & .002 & - & .049 & .025 & .059 & .040 & .100 & .222 & .113 & .041 & .041 & .003 & .015 \\
.072 & .015 & .004 & .014 & .060 & .008 & .005 & .282 & .033 & - & .138 & .012 & .013 & .021 & .031 & .030 & .053 & .113 & .006 & .025 \\
.069 & .019 & .012 & .020 & .084 & .022 & .014 & .135 & .070 & .566 & - & .026 & .033 & .127 & .070 & .101 & .050 & .239 & .015 & .026 \\
.008 & .004 & .115 & .069 & .010 & .088 & .053 & .018 & .076 & .022 & .012 & - & .018 & .055 & .075 & .191 & .082 & .012 & .001 & .014 \\
.065 & .005 & .041 & .046 & .005 & .031 & .011 & .012 & .048 & .024 & .014 & .017 & - & .034 & .022 & .057 & .047 & .037 & .002 & .009 \\
.067 & .004 & .065 & .307 & .006 & .030 & .035 & **.000** & .153 & .048 & .070 & .066 & .043 & - & .153 & .124 & .046 & .040 & .007 & .037 \\
.085 & .005 & .015 & .086 & .013 & .036 & .038 & .020 & .282 & .058 & .032 & .075 & .023 & .127 & - & .050 & .059 & .003 & .004 & .017 \\
.233 & .015 & .074 & .087 & .024 & .094 & .018 & .036 & .092 & .036 & .030 & .121 & .039 & .066 & .032 & - & .164 & .005 & .003 & .014 \\
.113 & .015 & .043 & .046 & .021 & .024 & .009 & .030 & .043 & .081 & .019 & .067 & .041 & .032 & .049 & .212 & - & .126 & .007 & .014 \\
.161 & .014 & .012 & .043 & .022 & .022 & .005 & .595 & .040 & .161 & .083 & .009 & .029 & .025 & .002 & .006 & .116 & - & .003 & .031 \\
.020 & .009 & .008 & .018 & .077 & .037 & .008 & .009 & .017 & .043 & .026 & .004 & .008 & .024 & .016 & .018 & .033 & .014 & - & .094 \\
.037 & .009 & .012 & .029 & .241 & .019 & .088 & .040 & .030 & .071 & .018 & .021 & .015 & .047 & .026 & .034 & .027 & .063 & .037 & -
\end{pmatrix}$$

Figure 2

Regularized rate matrix generated from BLOSUM62 Regularized rate matrix generated from BLOSUM62 after the QOG algorithm has been applied. The matrix has been rescaled to have an average number of one substitution per amino acid. In this simple case in which there was at most one negative off-diagonal entry per row, the regularization process requires the negative off-diagonal value to be set to zero (represented in bold in this Figure), and to shift the rest of the elements in that row by the corresponding amount so that the sum of all elements is zero. Rows without any off-diagonal negative values remain unchanged from the values obtained in Figure 1.

the QOG algorithm we obtain a regularized version of the rate matrix using BLOSUM62, which is given in Figure 2.

Regularization algorithms

Here I reproduce the QOG and QOM regularization algorithms. The proofs for these algorithms can be found in [57]. The QOG algorithm regularizes each row of a rate matrix independently. Given a row in a rate matrix R,

$$r = (r_1,...,r_n) \equiv (R(i, 1) ..., R(i, n)), \quad (5)$$

the QOG algorithm solves the problem of finding the vector at the minimal Euclidean distance from r such that the sum of all its elements is zero, and all elements but one are positive.

The steps of the QOG algorithm are:

1. Permute the row vector so that $r_1 = R(i, i)$.

2. Construct the vector w, such that $w_i = r_i - \lambda$, where $\lambda = \frac{1}{n}\sum_{i=1}^{n} r_i$.

3. Obtain the permutation $w^p = P(w)$, such that $w_i^p \le w_{i+1}^p$.

4. Construct $C_k = w_1^p + \sum_{i=0}^{n-k-1} w_{n-i}^p - (n-k+1)w_{k+1}^p$, for $k = 2, ..., n - 1$.

5. Calculate $k_{min} = \min_k = 2, ..., n - 1 \{k \text{ such that } C_k \le 0\}$.

6. Construct the vector

$$\hat{r}_i = \begin{cases} 0 & \text{if } 2 \le i \le k_{min} \\ w_i^p - \frac{1}{n-k_{min}+1}\left\{ w_1^p + \sum_{j=k_{min}+1}^{n} w_j^p \right\} & \text{otherwise} \end{cases}$$

7. The regularized row is given by $r \leftarrow P^{-1}(\hat{r})$. Finally reverse the permutation of step (1).

The QOM algorithm regularizes each row of a conditional matrix independently. Given a row in a conditional matrix Q_t

$$r = (r_1,..., r_n) \equiv (Q_t(i, 1),..., Q_t(i, n)), \quad (6)$$

the QOM algorithm solves the problem of finding the vector at the minimal Euclidean distance from r such that the sum of all its elements is one, and all elements are positive.

The steps of the QOM algorithm are:

1. Construct the vector w, such that $w_i = r_i - \lambda$, where $\lambda = \frac{1}{n}(\sum_{i=1}^{n} r_i - 1)$.

2. If all w_i are non negative, $r \leftarrow w$ is the new regularized row.

3. Otherwise, obtain the permutation $w^p = P(w)$, such that $w_i^p \geq w_{i+1}^p$.

4. Construct $C_k = \sum_{i=1}^{k} w_i^p - k w_k^p$, for $k = 1,..., n$.

5. Calculate $k_{max} = \max_{k=1,...,n}\{k \text{ such that } C_k \leq 1\}$.

6. Construct the vector

$$\hat{r}_i = \begin{cases} w_i^p + \frac{1}{k_{max}}[1 - \sum_{j=1}^{k_{max}} w_j^p] & \text{if } 1 \leq i \leq k_{max} \\ 0 & \text{otherwise} \end{cases}$$

7. The regularized row is given by $r \leftarrow P^{-1}(\hat{r})$.

A 4 × 4 example starting from joint probabilities at a given generating time

As an review of these techniques, I will use a set of 4 × 4 single-nucleotide joint probabilities $P(i, j|t_*)$ for $i, j = \{a, c, g, t\}$ at a particular generating time t_* to construct the corresponding rate matrix.

In this example, the joint probabilities at the generating time using the matrix notation $P_*(ij) \equiv P(i, j|t_*)$ are given by,

$$P_* = \begin{pmatrix} A & C & G & T \\ 0.1248 & 0.0520 & 0.0611 & 0.0457 \\ 0.0520 & 0.0873 & 0.0448 & 0.0470 \\ 0.0611 & 0.0448 & 0.1087 & 0.0385 \\ 0.0457 & 0.0470 & 0.0385 & 0.1010 \end{pmatrix}. \quad (7)$$

These 4 × 4 pair-nucleotide probabilities are taken from the program QRNA. They were calculated according to [16] by marginalizing codon-codon joint probabilities which were constructed from the BLOSUM62 matrix of amino acid substitutions. These 4 × 4 probabilities can be viewed as a particular example of the REV model [44]. Note that the sum of all elements of P_* adds up to one, and the matrix is symmetric.

The marginal probabilities defined as $p_i = \sum_j P(i, j|t_*)$ can be calculated from the joint probabilities to be,

$$p = (p_a, p_c, p_g, p_t) = (0.2836, 0.2311, 0.2531, 0.2322). \quad (8)$$

Similarly, the conditional probabilities $P(j|i, t_*)$ can be calculated from the previous joint and marginal probabilities using the relationship $P(i, j|t_*) = P(j|i, t_*)\, p_i$. Using the matrix representation $Q_*(ij) \equiv P(j|i, t_*)$ we have,

$$Q_* = \begin{pmatrix} 0.4401 & 0.1834 & 0.2154 & 0.1611 \\ 0.2250 & 0.3778 & 0.1939 & 0.2034 \\ 0.2414 & 0.1770 & 0.4295 & 0.1521 \\ 0.1968 & 0.2024 & 0.1658 & 0.4350 \end{pmatrix}. \quad (9)$$

Notice how the sum of the elements in each row adds up to one. Notice also how Q_* is quite different from the identity matrix, which means that we have started with a quite divergent generating time.

If we assume a homogeneous Markov substitution process, we can interpret the conditional probabilities Q_* as the matrix of substitution probabilities at the generating time. Thus, we can characterize the underlying evolutionary process by its instantaneous rate of evolution, which can be calculated from Q_* using equation (4). The resulting rate matrix R (up to an arbitrary scaling factor t_*) is given by,

$$R = \frac{1}{t_*}\begin{pmatrix} -1.0965 & 0.3690 & 0.4575 & 0.2701 \\ 0.4528 & -1.2750 & 0.3720 & 0.4503 \\ 0.5126 & 0.3396 & -1.0876 & 0.2353 \\ 0.3298 & 0.4481 & 0.2565 & -1.0345 \end{pmatrix}. \quad (10)$$

This rate matrix has all the good properties: (*i*)"Normalization": the sum of the elements of each row is zero. (*ii*) "Reversibility": $p_i R_{ij} = p_j R_{ji}$. The process is reversible by construction because we started with symmetric joint probabilities. (*iii*) "Saturation". The rate matrix converges at time infinity to the given marginal probabilities in equation (8). We can test saturation by using equation (2) and calculating the substitution matrix for a very large time. For instance, for $t = 10t_*$ we have

$$Q_{10.0t_*} = \begin{pmatrix} 0.2836 & 0.2311 & 0.2531 & 0.2322 \\ 0.2836 & 0.2311 & 0.2531 & 0.2322 \\ 0.2836 & 0.2311 & 0.2531 & 0.2322 \\ 0.2836 & 0.2311 & 0.2531 & 0.2322 \end{pmatrix}. \quad (11)$$

Saturation (or stationarity) of a Markov process is a necessary consequence of (*i*) normalization and (*ii*) reversibility. Appendix A shows a derivation of the previous statement which was useful for me (and hopefully for some readers) when studying the behavior at $t = \infty$ of more complicated evolutionary models. Therefore, starting from joint probabilities as in this example, we can always interpret the marginal probabilities as the stationary probabilities of the evolutionary process.

In summary, starting with a single set of joint probabilities at one particular generating divergence time t_*, we calculate the joint probabilities at any other arbitrary time, assuming an exponential model of evolution. To that effect, given the particular set of joint probabilities (7) we have calculated the corresponding rate matrix (10) by Taylor expansion. Thus we can estimate the substitution matrix/conditional probabilities at any other arbitrary time, simply using equation (2), and reconstruct the joint probabilities at any other arbitrary time. For instance, for $t = 0.3t_*$ we obtain,

$$P_{0.3t_*} = \begin{pmatrix} 0.2089 & 0.0249 & 0.0303 & 0.0195 \\ 0.0249 & 0.1615 & 0.0208 & 0.0239 \\ 0.0303 & 0.0208 & 0.1864 & 0.0156 \\ 0.0195 & 0.0239 & 0.0156 & 0.1731 \end{pmatrix}. \quad (12)$$

This method allows us to evolve pair emission probabilities corresponding to different processes (in addition to the 4 × 4 nucleotide emissions) for instance 20 × 20 amino acid-to-amino acid joint emission probabilities, 64 × 64 codon-to-codon joint emission probabilities, or 16 × 16 RNA basepair-to-basepair joint emission probabilities. Thus, this method is useful to be applied in combination with pair HMMs or pair SCFGs already parameterized at one fixed divergence time to make their emission probabilities a time-dependent family.

The evolution of emission probabilities with indels treated as an extra character

Substitution processes (even if describing multi-nucleotide events such as codon evolution or RNA basepair evolution) are not enough to describe the full evolutionary relationship between two biological sequences. We also need to consider indels, for which we need to introduce more complicated models of evolution than the one described so far.

Indels have traditionally been a problem for phylogenetic methods. Programs to construct phylogenetic trees from data such as PHYLIP [58], PAUP* [59], and other phylogeny packages [60-64] treat gaps as missing data. The theoretical description of the evolution of gaps in a probabilistic fashion reached a landmark with the Thorne/Kishino/Felsenstein (TKF) model [30,31]. The TKF model however is hard to implement in combination with a probabilistic model such as an HMM, although an active area of research exists in that direction [36,39,40]. A more direct attack to the problem of introducing phylogeny into existing probabilistic models originated with the concept of tree HMMs [34,35]. The tree HMM method models the evolution of the parsing of different sequences through an HMM. This approach is more related with the evolution of transition probabilities, and I will discuss it later on in this paper.

Here I am going to describe a method for the evolution of indels under the assumption that they behave like an additional residue added to a $N \times N$ residue substitution matrix. This is a simplification of the problem because it forces indels to have linear penalties (that is, the cost of opening an indel in an alignment or the cost of extending it with one more indel character is the same) and to behave independently of each other (that is, successive indel characters in one sequence will be treated as independent events, rather than as a single indel of n residues long). Despite its apparent simplicity, this approach poses interesting problems in parameterizing evolution.

Let us review some of the implications of insertion and deletion processes. The treatment that pair models give to pairwise alignments can be interpreted (if we assume reversibility, as is the case here) with all generality as if one of the sequences is the ancestor of the other one. For any two aligned residues we assume that they can be related by a substitution process. For a residue aligned to a gap we assume that either a residue in the ancestor was deleted in the descendent sequence, or that a residue not present in the ancestor appeared in the descendent sequence.

An stochastic insertion–deletion process also involves insertions followed by subsequent deletions. These events leave no trace in pairwise alignments because alignments usually do not retain gaps aligned to gaps. However, when we are treating indels as an extra character, we have to account for such events.

If we were given ideal alignments with all their gap-to-gap aligned columns we could estimate from data the $(N + 1) \times (N + 1)$ extended joint probabilities at a generating time, P_*^{ε}. Because that is not the case, we need to make some inference about P_*^{ε}. Let us represent with Δ, such that $0 \leq \Delta \leq 1/2$, the expected frequency of observed gaps with respect to the total number of residues in pairwise alignments at a particular time t_*. The parameter Δ, can be estimated from data, or it could be estimated according to the TKF model [30] as

$$\Delta = \frac{\lambda}{2}\frac{1-e^{(\lambda-\mu)t_*}}{\mu-\lambda e^{(\lambda-\mu)t_*}}, \qquad (13)$$

if we knew the values for the rate of insertions λ and the rate of deletions μ, such that $0 < \lambda < \mu$.

Let us represent with Δ' the expected frequency of missing gap-to-gap aligned columns in a pairwise alignment at a particular time t_*. One can estimate Δ' as the expected length of insertions that were later deleted without leaving any trace in current sequences. The probability of a stretch of l gap-to-gap characters is given by the geometric distribution density $\rho(l) = (1 - \Delta^2)\,\Delta^{2l}$. Therefore Δ' is given by,

$$\Delta' = \sum_{l=1}^{\infty} l\rho(l) = \frac{\Delta^2}{1-\Delta^2}. \qquad (14)$$

Using these two parameters and the joint probabilities in the absence of gaps at the generating time $P_*(\hat{\imath}\hat{\jmath})$ we can construct the set of $(N + 1) \times (N + 1)$ extended joint probabilities at t_* as

$$\Omega\left(\frac{P_*(\imath\jmath) \mid p_\imath\Delta}{p_\jmath\Delta \mid \Delta'}\right), \qquad (15)$$

where we have assumed independence for the joint probability of a residue and a gap. The normalization factor $\Omega = 1/(1 + 2\Delta + \Delta')$ represents the fact that the observed Δ is different from the value we would have obtained had we known the complete alignment.

Another implication of insertion and deletions appears in the behavior of the marginal probabilities of single residues and indels. At $t = 0$ when sequences have not yet diverged, the marginal probability of finding a gap in an alignment should be zero. In the limit $t = \infty$, the pairwise alignment of two finite-length sequences is going to be dominated by gap-to-gap alignments, which implies that as the divergence time increases the marginal probability of a residue becomes negligible, while the marginal probability of a gap becomes one in the limit $t = \infty$. Our evolutionary model has to be able to accommodate such saturation frequencies.

A step-by-step description of the algorithm for the evolution of gaps as an extra residue

I will start by describing the steps to implement the method before explaining how to derive those steps. This method can be applied starting from two different situations: starting from a $N \times N$ set of joint probabilities at a generating time that need to be extended to allow indel characters and evolved with time; or starting from a given $N \times N$ rate matrix that needs to be extended to allow indel characters.

Suppose we start with a $N \times N$ set of joint probabilities P_* at a generating time t_*, where p stands for the marginal probabilities and Q_* represents the set of conditional probabilities associated with P_*.

1. Extend the joint probabilities at the generating time t_* to a $(N + 1) \times (N + 1)$ matrix of joint probabilities P_*^{ε} of the form,

$$P_*^{\varepsilon} = \Omega\left(\frac{P_*(\imath\jmath) \mid p_\imath\Delta}{p_\jmath\Delta \mid \Delta'}\right), \qquad (16)$$

where Δ is a parameter which represents the expected frequency of gaps with respect to the total number of residues in an pairwise alignment at t_*, and which satisfies the condition $0 \le \Delta \le 1/2$. The parameter Δ' is given in terms of Δ as $\Delta' = \frac{\Delta^2}{1-\Delta^2}$, and the normalization constant is given by $\Omega = 1/(1 + 2\Delta + \Delta')$. The indices with hats $(\hat{\imath})$ stand for the N residues, and exclude the gap character, which I represent with the symbol -.

The $(N + 1) \times (N + 1)$ extended conditional probabilities at the generating time Q_*^{ε} are given by,

$$Q_*^{\varepsilon} = \left(\begin{array}{c|c} Q_*(\imath\jmath)/(1+\Delta) & \begin{array}{c}\frac{\Delta}{1+\Delta}\\ \vdots \\ \frac{\Delta}{1+\Delta}\end{array} \\ \hline p_\jmath\frac{\Delta}{1+\Delta'} & \frac{\Delta'}{\Delta+\Delta'} \end{array}\right). \qquad (17)$$

2. Construct the $(N + 1) \times (N + 1)$ extended rate matrix R^{ε} as

$$R^{\varepsilon} = \frac{1}{t_*}\log(Q_0^{-1}Q_*^{\varepsilon}) = \frac{1}{t_*}\sum_{n=1}^{n=\infty}\frac{(-1)^{n+1}}{n}(Q_0^{-1}Q_*^{\varepsilon} - I)^n, \qquad (18)$$

where

$$Q_0^{-1}Q_*^{\varepsilon} = \left(\begin{array}{c|c} Q_*(\imath\jmath)/(1+\Delta) & \begin{array}{c}\frac{\Delta}{1+\Delta}\\ \vdots \\ \frac{\Delta}{1+\Delta}\end{array} \\ \hline 0 \quad \dots \quad 0 & 1 \end{array}\right). \qquad (19)$$

3. Calculate the exponential of the rate matrix $e^{tR^{\varepsilon}}$ using the Taylor expansion,

$$e^{tR^{\varepsilon}} = \sum_{n=0}^{n=\infty} \frac{(tR^{\varepsilon})^n}{n!}. \qquad (20)$$

4. Construct the extended matrix of conditional probabilities at arbitrary time Q_t^{ε} as

$$Q_t^{\varepsilon} = Q_0 e^{tR^{\varepsilon}}, \qquad (21)$$

where the matrix Q_0 is given by

$$Q_0 = \left(\begin{array}{c|c} \delta(\imath\jmath) & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ \hline p_{\jmath}(1-q_0) & q_0 \end{array} \right), \qquad (22)$$

where $p_{\hat{\imath}}$ are the original marginal probabilities of P_*, and the probability $0 < q_0 \leq 1$ is given by $q_0 = \frac{\Delta' - \Delta^2}{\Delta' + \Delta}$. The function $\delta(ij)$ is a Kronecker delta which takes value one for $i = j$ and zero otherwise. The case $q_0 = 1$ corresponds to the extreme case in which the $N + 1$ gap residue does not evolve.

5. Construct the extended marginal probabilities p_t^{ε} as

$$p_t^{\varepsilon}(i) = \begin{cases} p_1(1-\Lambda_t) & \text{if} \quad i = \imath, \\ \Lambda_t & \text{if} \quad i = -, \end{cases} \qquad (23)$$

where the probability of a gap at time t is given by,

$$\Lambda_t = \frac{\sum_{\hat{\imath}} p_{\hat{\imath}} Q_t^{\varepsilon}(\hat{\imath}-)}{\sum_{\hat{\imath}} p_{\hat{\imath}} Q_t^{\varepsilon}(\hat{\imath}-) + 1 - Q_t^{\varepsilon}(--)}. \qquad (24)$$

I call this process "quasi-stationary" because the background frequencies $p_t^{\varepsilon}(\hat{\imath})$ at any finite time are always proportional to the original N-dimensional background frequencies $p_{\hat{\imath}}$. This result is a concequence of the fact that the first N elements of the last row of Q_0 are proportional to the N stationary frequencies $p_{\hat{\imath}}$. On the other hand, while remaining "quasi-stationary" the background frequencies evolve from $(p_{\hat{\imath}}, 0)$ at time zero towards "all gaps" at time infinity, *i.e.* $\lim_{t\to\infty} \Lambda_t = 1$. This behaviour at time infinity is the consequence of the particular value of q_0 selected in the previous step.

6. Finally, construct the evolved $(N + 1) \times (N + 1)$ joint probabilities at arbitrary time P_t^{ε} as

$$P_t^{\varepsilon}(ij) = p_t^{\varepsilon}(i) Q_t^{\varepsilon}(ij). \qquad (25)$$

The expression for Λ_t in equation (24) guarantees reversibility, that is, that the extended P_t^{ε} constructed according to the above expression are symmetric.

For the other starting situation, in which we have a $N \times N$ rate marix R, the procedure to generate a $(N + 1) \times (N + 1)$ quasi-stationary reversible evolutionary model is the following:

1. Construct the $(N + 1) \times (N + 1)$ extended rate matrix R^{ε} as

$$R^{\varepsilon} = \left(\begin{array}{c|c} R(\imath\jmath) - \beta\delta(\imath\jmath) & \begin{matrix} \beta \\ \vdots \\ \beta \end{matrix} \\ \hline 0 \quad \ldots \quad 0 & 0 \end{array} \right), \qquad (26)$$

where we have extended the $N \times N$ rate matrix R with the parameter $\beta > 0$.

The instantaneous rate is given by,

$$Q_0 R^{\varepsilon} = \left(\begin{array}{c|c} R(\imath\jmath) - \beta\delta(\imath\jmath) & \begin{matrix} \beta \\ \vdots \\ \beta \end{matrix} \\ \hline -\beta(1-q_0)p_1 \quad \ldots \quad -\beta(1-q_0)p_N & \beta(1-q_0) \end{array} \right). \qquad (27)$$

Thus β is the instantaneous rate of deletion of a character, while $-\beta(1 - q_0)\, p_{\hat{\imath}}$ is the rate of insertion of character $\hat{\imath}$. (More complicated models in which the rate of deletion is different for different characters are also possible.) Notice that $q_0 = 1$ corresponds to the case in which the rate of insertions is zero.

2. Find $e^{tR^{\varepsilon}}$ analytically, if an analytic expression for R^{ε} is given by solving the differential equation $d\,(e^{tR^{\varepsilon}})\,/\,dt = R^{\varepsilon}\, e^{tR^{\varepsilon}}$, or numerically, proceeding as in step (3) of the previous procedure.

3. Proceed as in steps (4)-(6) of the previous procedure.

A 5 × 5 example starting from joint probabilities at a given generating time

We start with the generating joint probabilities P_* in the 4 × 4 example in equation (7), which we want to extend to a 5 × 5 matrix by adding a gap character. For this example, I have selected the arbitrary value for the gap parameter $\Delta = 0.18$.

The 4 × 4 joint probabilities in equation (7) augmented to a 5 × 5 matrix P_*^{ε} using the gap parameter Δ = 0.18 (which implies that Δ' = $\Delta^2/(1 - \Delta^2)$ = 0.0335) is given by,

$$P_*^{\varepsilon} = \begin{pmatrix} A & C & G & T & - \\ 0.0896 & 0.0373 & 0.0438 & 0.0328 & 0.0366 \\ 0.0373 & 0.0626 & 0.0321 & 0.0337 & 0.0299 \\ 0.0438 & 0.0321 & 0.0780 & 0.0276 & 0.0327 \\ 0.0328 & 0.0337 & 0.0276 & 0.0725 & 0.0300 \\ 0.0366 & 0.0299 & 0.0327 & 0.0300 & 0.0240 \end{pmatrix}. \quad (28)$$

The conditional probabilities $Q_*^{\varepsilon}(ij) = P^{\varepsilon}(j|i, t_*)$ are given by,

$$Q_*^{\varepsilon} = \begin{pmatrix} 0.3729 & 0.1554 & 0.1826 & 0.1366 & 0.1525 \\ 0.1907 & 0.3201 & 0.1643 & 0.1724 & 0.1525 \\ 0.2046 & 0.1500 & 0.3640 & 0.1289 & 0.1525 \\ 0.1668 & 0.1715 & 0.1405 & 0.3686 & 0.1525 \\ 0.2391 & 0.1949 & 0.2134 & 0.1958 & 0.1568 \end{pmatrix}. \quad (29)$$

The extended marginal probabilities at the particular time instance t_* are given by,

$$p_{t_*}^{\varepsilon} = (p_a, p_c, p_g, p_t, p_-) = (0.2402, 0.1957, 0.2143, 0.1966, 0.1532), \quad (30)$$

which are quasi-stationary with respect to the 4 × 4 stationary probabilities p = (0.2836, 0.2311, 0.2532, 0.2322) we started with in equation (8).

The matrix of conditional probabilities at time zero using expression (22) is given by,

$$Q_0 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0.2822 & 0.2299 & 0.2518 & 0.2310 & 0.0051 \end{pmatrix}. \quad (31)$$

The rate matrix for this example, calculated using the Taylor expansion described in equation (18) takes the value,

$$R^{\varepsilon} = \begin{pmatrix} -1.2625 & 0.3692 & 0.4578 & 0.2700 & 0.1655 \\ 0.4531 & -1.4415 & 0.3721 & 0.4508 & 0.1655 \\ 0.5130 & 0.3398 & -1.2535 & 0.2352 & 0.1655 \\ 0.3298 & 0.4486 & 0.2564 & -1.2003 & 0.1655 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \quad (32)$$

One should not be concerned to see a whole row of zeros for this rate matrix. For this generalized model the instantaneous rate of evolution is not directly given by the rate matrix; instead, the instantaneous rate of evolution is given by,

$$\left.\frac{dQ_t^{\varepsilon}}{dt}\right|_{t=0} = Q_0 R^{\varepsilon}. \quad (33)$$

In this example, the instantaneous rate of evolution takes the form,

$$Q_0 R^{\varepsilon} = \begin{pmatrix} -1.2625 & 0.3692 & 0.4578 & 0.2700 & 0.1655 \\ 0.4531 & -1.4415 & 0.3721 & 0.4508 & 0.1655 \\ 0.5130 & 0.3398 & -1.2535 & 0.2352 & 0.1655 \\ 0.3298 & 0.4486 & 0.2564 & -1.2003 & 0.1655 \\ -0.0467 & -0.0381 & -0.0417 & -0.0382 & 0.1647 \end{pmatrix}. \quad (34)$$

One should not be concerned either by having some negative off diagonal components. For small times δt, the conditional matrix is given by,

$$Q_{\delta t}^{\varepsilon} = Q_0 e^{\delta t R^{\varepsilon}} \approx Q_0 + \delta t (Q_0 R^{\varepsilon}). \quad (35)$$

Therefore, in order to have a proper matrix of conditional probabilities for sufficiently small δt, it is necessary to satisfy the following condition for each pair of indices i, j,

$$\text{if } Q_0(ij) = 0 \text{ then } (Q_0 R^{\varepsilon})(ij) > 0. \quad (36)$$

In this case, the off-diagonal components of the last row of Q_0 are non-zero, which allows us to have negative off-diagonal elements for that row in the instantaneous rate matrix $Q_0 R^{\varepsilon}$.

With the 5 × 5 rate matrix in hand, we can apply steps (3) and (4) to obtain the conditional probabilities at any arbitrary time Q_t^{ε}. For instance for $t = 0.3t_*$ we obtain the following evolved conditional probabilities:

$$Q_{0.3t_*}^{\varepsilon} = \begin{pmatrix} 0.7011 & 0.0834 & 0.1017 & 0.0654 & 0.0484 \\ 0.1023 & 0.6651 & 0.0856 & 0.0985 & 0.0484 \\ 0.1139 & 0.0781 & 0.7007 & 0.0588 & 0.0484 \\ 0.0799 & 0.0981 & 0.0641 & 0.7095 & 0.0484 \\ 0.2685 & 0.2188 & 0.2396 & 0.2198 & 0.0533 \end{pmatrix}. \quad (37)$$

The quasi-stationary marginal probabilities are constructed using the result $\Lambda_{0.3t_*} = 0.0487$, and the 4 × 4 stationary probabilities p = (0.2836, 0.2311, 0.2532, 0.2322), following step (5) of the algorithm as,

$$p_{0.3t_*}^{\varepsilon} = (0.2698, 0.2199, 0.2408, 0.2209, 0.0487). \quad (38)$$

Finally, using equation (25), for $t = 0.3t_*$, we obtain the following evolved joint probabilities

$$P^{\varepsilon}_{0.3t_*} = \begin{pmatrix} 0.1891 & 0.0225 & 0.0274 & 0.0176 & 0.0131 \\ 0.0225 & 0.1462 & 0.0188 & 0.0217 & 0.0106 \\ 0.0274 & 0.0188 & 0.1687 & 0.0142 & 0.0117 \\ 0.0176 & 0.0217 & 0.0142 & 0.1567 & 0.0107 \\ 0.0131 & 0.0106 & 0.0117 & 0.0107 & 0.0026 \end{pmatrix}. \quad (39)$$

Notice that this matrix is symmetric, which is the result of having imposed reversibility for any arbitrary divergence time.

We can also see by calculating the conditional probabilities at large divergence times how these probabilities evolve towards their saturation values given by (0, 0, 0, 0, 1). For instance, for $t = 30t_*$ we have,

$$Q^{\varepsilon}_{30t_*} = \begin{pmatrix} 0.0020 & 0.0016 & 0.0018 & 0.0016 & 0.9930 \\ 0.0020 & 0.0016 & 0.0018 & 0.0016 & 0.9930 \\ 0.0020 & 0.0016 & 0.0018 & 0.0016 & 0.9930 \\ 0.0020 & 0.0016 & 0.0018 & 0.0016 & 0.9930 \\ 0.0019 & 0.0016 & 0.0017 & 0.0016 & 0.9933 \end{pmatrix}. \quad (40)$$

An example starting from a rate matrix: The Jukes-Cantor model extended to gaps

As an example of a situation in which we start with a rate matrix, let us consider the generalization of the Jukes-Cantor model [42] to a 5 × 5 evolutionary model with a gap character. The original Jukes-Cantor model assumes that all nucleotides mutate at the same rate $\alpha > 0$ which is represented by the rate matrix

$$R = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}. \quad (41)$$

In this simple case the conditional matrix $Q_t = e^{tR}$ can be found analytically by solving the matrix differential equation $\dot{Q}_t = RQ_t$. Because of the symmetries of the problem we can write

$$Q_t = \begin{pmatrix} r_t & s_t & s_t & s_t \\ s_t & r_t & s_t & s_t \\ s_t & s_t & r_t & s_t \\ s_t & s_t & s_t & r_t \end{pmatrix}, \quad (42)$$

with the condition $r_t + 3s_t = 1$. We then obtain the following differential equations

$$\dot{r}_t = -3\alpha r_t + 3\alpha s_t, \quad (43)$$

$$\dot{s}_t = -\alpha s_t + \alpha r_t, \quad (44)$$

and the solutions are,

$$r_t = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}, \quad (45)$$

$$s_t = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}, \quad (46)$$

By taking the limit $t = \infty$ in the previous two equations, one can see that the saturation frequencies of the Jukes-Cantor model are $p_i = 0.25$ for $i = a, c, g, t$.

The 5 × 5 extended Jukes-Cantor rate matrix R^{ε} is constructed by adding a rate of mutation to a gap represented by the quantity $\beta \geq 0$ which in principle we will assume is different from the rate of substitutions α,

$$R^{\varepsilon} = \begin{pmatrix} -(3\alpha+\beta) & \alpha & \alpha & \alpha & \beta \\ \alpha & -(3\alpha+\beta) & \alpha & \alpha & \beta \\ \alpha & \alpha & -(3\alpha+\beta) & \alpha & \beta \\ \alpha & \alpha & \alpha & -(3\alpha+\beta) & \beta \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \quad (47)$$

We also introduce the matrix at time zero Q_0 which depends on the probability parameter $1 \geq q_0 > 0$,

$$Q_0 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ (1-q_0)/4 & (1-q_0)/4 & (1-q_0)/4 & (1-q_0)/4 & q_0 \end{pmatrix}, \quad (48)$$

where the particular case $q_0 = 1$ is only allowed if simultaneously $\beta = 0$, and corresponds to a trivial extension of the original Jukes-Cantor model in which the gap character does not evolve.

The conditional matrix at arbitrary time is given by $Q^{\varepsilon}_t = Q_0 e^{tR^{\varepsilon}}$. The symmetries of the problem in this case allow us to parameterize Q^{ε}_t as

$$Q^{\varepsilon}_t = \begin{pmatrix} r_t & s_t & s_t & s_t & \gamma_t \\ s_t & r_t & s_t & s_t & \gamma_t \\ s_t & s_t & r_t & s_t & \gamma_t \\ s_t & s_t & s_t & r_t & \gamma_t \\ \xi_t & \xi_t & \xi_t & \xi_t & \sigma_t \end{pmatrix}, \quad (49)$$

with the conditions $r_t + 3s_t + \gamma_t = 1$ and $4\xi_t + \sigma_t = 1$.

Introducing the matrix $M_t \equiv e^{tR^{\varepsilon}}$, we can parameterize

$$M_t = \begin{pmatrix} r_t & s_t & s_t & s_t & \gamma_t \\ s_t & r_t & s_t & s_t & \gamma_t \\ s_t & s_t & r_t & s_t & \gamma_t \\ s_t & s_t & s_t & r_t & \gamma_t \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad (50)$$

which implies that

$$\xi_t = \frac{1}{4}(1-q_0)(1-\gamma_t), \qquad (51)$$

$$\sigma_t = (1 - q_0)\,\gamma_t + q_0. \qquad (52)$$

The differential equation to calculate M_t takes the form $\dot{M}_t = R^{\varepsilon} M_t$, which translates into the differential equations,

$$\dot{r}_t = -(3\alpha+\beta)r_t + 3\alpha s_t, \qquad (53)$$

$$\dot{s}_t = -(\alpha+\beta)s_t + \alpha r_t, \qquad (54)$$

$$\dot{\gamma}_t = \beta(1-\gamma_t). \qquad (55)$$

Which are satisfied by

$$r_t = \frac{1}{4}e^{-\beta t} + \frac{3}{4}e^{-(4\alpha+\beta)t}, \qquad (56)$$

$$s_t = \frac{1}{4}e^{-\beta t} - \frac{1}{4}e^{-(4\alpha+\beta)t}, \qquad (57)$$

$$\gamma_t = 1 - e^{-\beta t}. \qquad (58)$$

And in addition we have

$$\xi_t = \frac{1}{4}(1-q_0)e^{-\beta t}, \qquad (59)$$

$$\sigma_t = 1 - (1 - q_0)\,e^{-\beta t}. \qquad (60)$$

In the limit case $\beta = 0$, the solutions for r_t and s_t reduce to those of the original Jukes-Cantor model with the trivial additions of $\sigma_t = 1$, $\xi_t = 0$ and $\gamma_t = 0$, after setting $q_0 = 1$.

The extended Jukes-Cantor model depends on three parameters: the rate of nucleotide substitution $\alpha > 0$, the rate of nucleotide deletion $\beta \geq 0$, and the parameter $1 \geq q_0 > 0$. What is the meaning of q_0 ? q_0 controls the saturation frequencies (*i.e.* the background frequencies at time infinity), as well as the background frequencies at any other finite time. For $\beta > 0$ and $1 > q_0 > 0$, taking the limit $t = \infty$ in equations (56)-(60), one can see that the saturation probabilities are given by (0, 0, 0, 0, 1). At any other finite time, the background frequencies of the model are quasi-stationary with respect to the background frequencies of the original Jukes-Cantor model, and are given by

$$p_t(i) = \frac{1}{4}(1-\Lambda_t), \qquad (61)$$

$$p_t(-) = \Lambda_t. \qquad (62)$$

Imposing the reversibility condition $p_t^T Q_t = p_t^T$ in particular we obtain $(1 - \Lambda_t)\gamma_t + \Lambda_t\sigma_t = \Lambda_t$ which implies,

$$\Lambda_t = \frac{1-e^{-\beta t}}{1-q_0 e^{-\beta t}}. \qquad (63)$$

Therefore q_0 controls how fast the background frequencies approach the saturation probabilities (0, 0, 0, 0, 1) through the factor Λ_t. For a given β, the larger q_0, the faster Λ_t approaches one. (Note that Λ_t always approaches one as t goes to infinity.)

At first glance, it looks like q_0 could take any value including one in the solution for the extended Jukes-Cantor model. $q_0 = 1$ would result in fixed background frequencies of the form (0, 0, 0, 0, 1), which is an undesirable result, and the value $q_0 = 1$ would have to be excluded when $\beta > 0$. In fact, the limit to the ungapped Jukes-Cantor model has to be taken by setting $\beta = 0$ first, and then $q_0 = 1$. In that way, $\Lambda_t = 0$ for all times, which is the correct result for the original Jukes-Cantor model.

Derivation of the algorithm for a (N + 1) × (N + 1) quasi-stationary and reversible evolutionary process

Unlike the ungapped $N \times N$ case in which the marginal probabilities are time independent, in the presence of gaps the marginal probabilities have to evolve with time. In fact, as I discussed earlier, the marginal probability of a gap $p_t^{\varepsilon}(-)$ has to evolve from zero at time zero to one at time infinity. As a result of that observation, probabilistic evolutionary models with $Q_0 \neq I$ are necessary in the presence of gaps in order to maintain reversibility. The reason for this requirement is the following: for an evolutionary model of the form e^{tR}, reversibility implies that there is some p_* such that $p_* Q_* = p_*$ [see Appendix B, equation (202)]; it follows then that $p_* R = 0$, and therefore $p_* e^{tR} = p_*$ for arbitrary time t. Thus, under a reversible model of the form $Q_t = e^{tR}$, marginal probabilities do not evolve with time. On the other hand, if $Q_0 \neq I$ then the condition $p_* Q_* = p_*$ does not imply $p_* R = 0$ for the rate matrix R, and therefore it does not impose p_* as the marginal probabilities for arbitrary t. (See appendix B for more details on this point.)

Therefore, to model the evolution of gaps we need to generalize the evolutionary model to have the following form

$$Q_t^{\varepsilon} = Q_0 e^{tR^{\varepsilon}}. \quad (64)$$

The matrix Q_0 can be parameterized in following form,

$$Q_0 = \left(\begin{array}{c|c} \delta(ij) & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ \hline p_j(1-q_0) & q_0 \end{array} \right), \quad (65)$$

This matrix depends on one additional parameter q_0. The particular dependency $Q_0(\text{-}, \hat{\imath}) \propto p_i$, is necessary to obtain quasi-stationary reversibility of the marginal probablilities.

The rate matrix is now a function of Q_0 and $Q_*^{\varepsilon} \equiv Q_{t=t_*}^{\varepsilon}$, and takes the form

$$R^{\varepsilon} = \frac{1}{t_*} \log(Q_0^{-1} Q_*^{\varepsilon}), \quad (66)$$

where the matrix $Q_0^{-1} Q_*^{\varepsilon}$ has the form,

$$Q_0^{-1} Q_*^{\varepsilon} = \left(\begin{array}{c|c} Q_*(ij)/(1+\Delta) & \begin{matrix} \frac{\Delta}{1+\Delta} \\ \vdots \\ \frac{\Delta}{1+\Delta} \end{matrix} \\ \hline p_j(1-\eta) & \eta \end{array} \right), \quad (67)$$

where

$$\eta = \left(1 - \frac{1}{q_0}\right) \frac{\Delta}{1+\Delta} + \frac{1}{q_0} \frac{\Delta}{\Delta + \Delta'}. \quad (68)$$

Notice that Q_0 may be inverted as long as $0 < q_0 \leq 1$.

With respect to the marginal probabilities we have that at the generating time t_* because of the way the extended probabilities P_* were constructed we imposed a quasi-stationary behavior of the form,

$$p_{t_*}^{\varepsilon} = [p_i(1 - \Lambda_{t_*}), \Lambda_{t_*}], \quad (69)$$

where

$$\Lambda_{t_*} = \frac{\Delta + \Delta'}{1 + 2\Delta + \Delta'}. \quad (70)$$

The generalized conditional matrix in (64) also saturates at very large times, and the saturation probabilities (*i.e.* the marginal probabilities at infinity) are given by those of the rate matrix, that is $\lim_{t\to\infty} Q_t^{\varepsilon} = \lim_{t\to\infty} e^{tR^{\varepsilon}}$ (see Corollary A.1). Because of the relationship in equation (66) between the rate matrix and the matrix $Q_0^{-1} Q_*^{\varepsilon}$, the saturation probabilities p_{∞}^{ε} are given by the condition (see Appendix B),

$$p_{\infty}^{\varepsilon}(i)(Q_0^{-1} Q_*^{\varepsilon})(ij) = p_{\infty}^{\varepsilon}(j)(Q_0^{-1} Q_*^{\varepsilon})(ji). \quad (71)$$

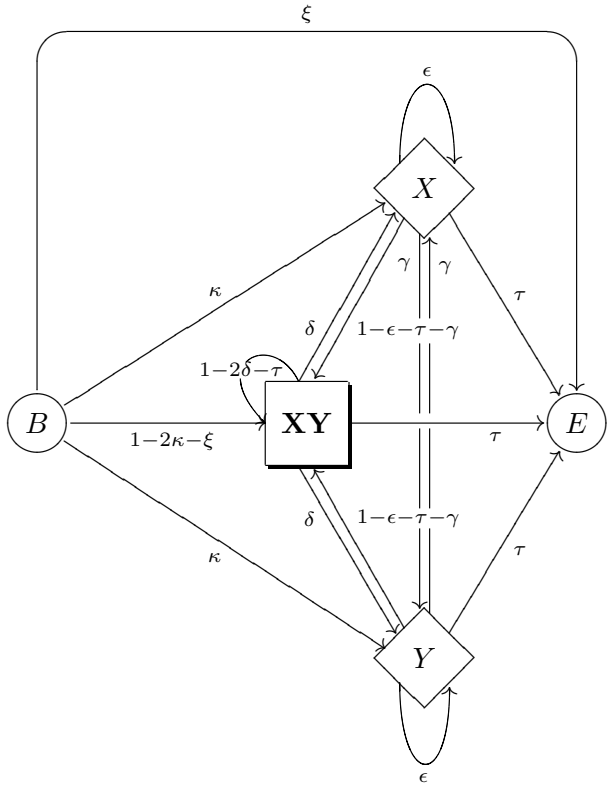

Figure 3
A pair-HMM model. Description of a pair-HMM model. The three states: Emit-a-pair (XY), Emit-X (X), and Emit-Y (Y) have four possible transitions each, which we are going to make time-dependent functions. This a geometric model, in which the expected length is given by $1/\tau$. In order to generate alignments with the same expected length at all times, we will leave the parameter τ (the transition of each of the three states into the exit state) unchanged with time. The figure shows the transition probabilities having the following properties: $T_t^{XY\to X} = T_t^{XY\to X}$, $T_t^{X\to X} = T_t^{Y\to Y}$, $T_t^{X\to Y} = T_t^{Y\to X}$, $T_t^{B\to X} = T_t^{B\to Y}$. These properties guarantee that the model is reversible.

Then using equation (71) we can see that the saturation probabilities maintain the quasi-stationary property that was imposed at the time instance t_*, and are given by,

$$p_\infty^\varepsilon = [p_i(1-\Lambda_\infty), \Lambda_\infty], \qquad (72)$$

where

$$\Lambda_\infty = \frac{\Delta}{\Delta + (1+\Delta)(1-\eta)}. \qquad (73)$$

As I discussed before, it is reasonable to impose that at infinity all we find is gaps, *i.e* $\Lambda_\infty = 1.0$, which implies $\eta = 1$ and

$$q_0 = \frac{\Delta' - \Delta^2}{\Delta' + \Delta}. \qquad (74)$$

Notice that because the relationship given in equation (14) between Δ' and Δ, then $0 < q_0 < 1$.

For an arbitrary time we have the reversibility relationship

$$p_t^\varepsilon(i)(Q_0^{-1}Q_t^\varepsilon)(ij) = p_t^\varepsilon(j)(Q_0^{-1}Q_t^\varepsilon)(ji). \qquad (75)$$

This equation is satisfied by construction in the $N \times N$ subspace. By inspecting the implications of the above equation for the gap index, we obtain an expression of Λ_t (the marginal probability of a gap) at arbitrary time that allows us to have quasi-stationary reversible evolution. The function Λ_t is given by,

$$\Lambda_t = \frac{\sum_{\hat{i}} p_{\hat{i}} Q_t^\varepsilon(\hat{i}-)}{\sum_{\hat{i}} p_{\hat{i}} Q_t^\varepsilon(\hat{i}-) + 1 - Q_t^\varepsilon(--)}. \qquad (76)$$

Evolutionary model for transition probabilities

The standard way in which comparative probabilistic models allow for insertions and deletions is by introducing several additional states with their corresponding transition probabilities. For instance, in a pair-HMM for sequence alignment (Figure 3) the presence of gaps requires the introduction of two states ("X" and "Y") which emit a nucleotide in only one of the two sequences. The probabilities associated with transitioning in and out of those states control the "gappiness" of the alignment. Therefore the evolution of these parameters with time is necessary in order to model different degrees of sequence divergence.

There has been a continued effort on improving the accuracy of the evolution of emission probabilities (*i.e.* substitution matrices) such as allowing correlations between the rates at different sites [65,66], improvements in the derivation of rate matrices from sequence data [23,67], or estimating multiple nucleotide changes [68]. In comparison, the ideas to describe the evolution of transition parameters in probabilistic models are much less standardized [34-40].

The goal of this section is to describe the evolution of transition probabilities. For instance, in the pair-HMM of Figure 3 the transition probabilities from the "**XY**" state to the "X" or "Y" states describe the introduction of gaps in one of the two aligned sequences, using an affine penalty. These transitions should be zero when the sequences have not yet diverged (time zero), but they should be maximal at infinite divergence. In between these two extremes, it is desirable to model the transition probabilities changing with divergence time. These methods are termed "evolutionary" because the transition probabilities will be parameterized with time, using functions that are generalizations of the Markov process that probabilistic evolutionary models assume for substitutions. Unlike the TKF model [30,31] and other related evolutionary models [32,33,41], the approach presented here will not describe the actual underlying evolutionary process that may have generated one sequence from another.

The tree-HMM method [34,35,37] is possibly the method closest to what I develop here. A tree HMM tries to model the phylogenetic relationship between related sequences by modeling the parsings of different sequences through the model. In a tree HMM it is not the actual transition probabilities of the HMM, but the parsing of the different sequences through the models that are evolved using rate matrices that resemble the diagonal rate matrices introduced in the first of the methods described below. Here I want to generate pair or profile probabilistic models that when comparing two related sequences are able to accommodate to the degree of divergence observed between the two sequences, and I intend to do that in a continuous-in-time and probabilistic fashion, using the smallest possible number of free parameters. No evolutionary history of individual insertion/deletion events will be generated; only *a posteriori* would an evolutionary history be established by comparing sequences (in the case of a profile model) or alignments (in the case of pair models) generated by the model at different times.

I present two methods to evolve transition probabilities. One of the methods considers the evolution of a vector of transition probabilities. In this method, the value of the transition probabilities at time zero and time infinity are input parameters, which gives a relatively large number of free parameters. In the second, more restrictive, method the transitions associated with several states are assumed to evolve under the same evolutionary process. This condition constrains some of the free parameters, but does

not fix them all completely. When the more restrictive conditions are used, both algorithms give the same results. These two algorithms are applicable to most pair and profile probabilistic models, be they HMMs or SCFGs, generalized or not. I present an example of the evolution of a vector probability vector for a pair HMM, and an example of the evolution of a matrix of transition substitutions for a profile HMM.

Evolution of a vector of transition probabilities

A step-by-step description of the algorithm

Let us start by providing the recipe to apply the algorithm:

1. Given a transition probability vector

$$q_t = \begin{pmatrix} T_t^1 \\ \vdots \\ T_t^n \end{pmatrix} \qquad \text{such that } \sum_{i=1}^{n} T_t^i = 1, \ \forall t, \tag{77}$$

2. Assume its set of values is known at the three particular time instances of $t = 0$, $t = t_*$, and $t = \infty$, named q_0, q_*, and q_∞. Assume each component i in these probability vectors satisfies one of the following three conditions,

$q_0(i) < q_*(i) < q_\infty(i)$ or $q_0(i) > q_*(i) > q_\infty(i)$ or $q_0(i) = q_*(i) = q_\infty(i)$, for all i. (78)

3. If the three input vectors satisfy the condition,

$$\frac{q_*(i) - q_\infty(i)}{q_0(i) - q_\infty(i)} = -r, \tag{79}$$

where $r > 0$ is a real number independent of i, then calculate q_t at an arbitrary time t ($0 < t < t_\infty$) as

$$q_t^T = q_\infty^T + (q_0 - q_\infty)^T \begin{pmatrix} e^{-rt/t_*} & & \\ & \ddots & \\ & & e^{-rt/t_*} \end{pmatrix}. \tag{80}$$

Normalization of the vector q_t is guaranteed by equation (79).

4. Otherwise q_t is given by the following expression

$$q_t^T = \frac{q_\infty^T}{1 + w_t} + \frac{(q_0 - q_\infty)^T}{1 + w_t} \begin{pmatrix} \exp[t/t_* \log(\frac{q_*(1) - q_\infty(1)}{q_0(1) - q_\infty(1)})] & & \\ & \ddots & \\ & & \exp[t/t_* \log(\frac{q_*(n) - q_\infty(n)}{q_0(n) - q_\infty(n)})] \end{pmatrix}, \tag{81}$$

where the function w_t is given by

$$w_t = \sum_{i=1}^{n} [q_0(i) - q_\infty(i)] \exp\left[t/t_* \log\left(\frac{q_*(i) - q_\infty(i)}{q_0(i) - q_\infty(i)} \right) \right]. \tag{82}$$

An example: evolution of the transition probabilities of a pair-HMM "XY" state

Consider the transition probability vector associated to the "XY" state of the pair-HMM given in Figure 3,

$$q_t^{XY} = \begin{pmatrix} T_t^{XY \to XY} \\ T_t^{XY \to X} \\ T_t^{XY \to Y} \\ T_t^{XY \to E} \end{pmatrix}, \tag{83}$$

which describe the four possible transitions from a correlated emission of two nucleotides to another correlated emission in both sequences ($T_t^{XY \to XY}$); to a gap in sequence Y ($T_t^{XY \to X}$); to a gap in sequence X ($T_t^{XY \to Y}$); or to end the alignment ($T_t^{XY \to E}$).

Below are some arbitrary values for the transition vector at divergence times: $t = 0$, $t = t_*$, and $t = \infty$ associated with state "XY", q^{XY},

$$q_0^{XY} = \begin{pmatrix} 1.0(1-\tau) \\ 0.0(1-\tau) \\ 0.0(1-\tau) \\ \tau \end{pmatrix}, q_{t_*}^{XY} = \begin{pmatrix} 0.74(1-\tau) \\ 0.13(1-\tau) \\ 0.13(1-\tau) \\ \tau \end{pmatrix}, q_\infty^{XY} = \begin{pmatrix} 0.00(1-\tau) \\ 0.50(1-\tau) \\ 0.50(1-\tau) \\ \tau \end{pmatrix}. \tag{84}$$

The transition $T^{XY \to E} = \tau$ is related to the expected length of the alignments generated using the model. We typically want to keep that transition invariant through time, and correlated with the alignment length L: $\tau = 1/L$. (This pair HMM produces sequences with a geometric length distribution of mean $1/\tau$.) The other three transitions change with time from a situation of no gaps at time zero, to a situation at time infinity in which all there is present is gaps, because no residue in either sequence has a homologous residue in the other.

Transition probabilities at $t = 0$ and $t = \infty$ can be stated from first principles. Transitions at the generating time t_*, are estimated from data, at the same time that emission probabilities are estimated. The transition probabilities at any other time are given by applying the algorithm. Using equation (81) we obtain,

$$q_{0.3t^*}^{XY} = \begin{pmatrix} 0.8942(1-1/L) \\ 0.0529(1-1/L) \\ 0.0529(1-1/L) \\ 1/L \end{pmatrix}, q_{1.6t^*}^{XY} = \begin{pmatrix} 0.6550(1-1/L) \\ 0.1725(1-1/L) \\ 0.1725(1-1/L) \\ 1/L \end{pmatrix}, \tag{85}$$

Similarly to this "XY" state case, all the other transition probabilities that appear in the pair model of Figure 3 could be continuously parameterized with the divergence time of the alignment being scored. This algorithm can be

applied to any full set of transition probabilities emerging from a particular state in a given probabilistic model that must evolve with time.

Connection with a tree-HMM 2×2 match-transition matrix

In the original representation of a tree-HMM [35] the idea of a match-transition matrix is introduced. If one parse through the HMM generated a Match to Match (*MM*) transition, while another parse through the model generated a Match-to-Delete (*MD*) transition, one can consider the substitution of *MM* by *MD* similarly to a substitution of residues by the conditional probability $P(MD|MM, t)$. This leads to the concept of a 2 × 2 match-transition matrix given by,

$$\begin{array}{c} \\ MM \\ MD \end{array} \begin{array}{c} \begin{array}{cc} MM & MD \end{array} \\ \begin{pmatrix} P(MM \mid MM,t) & P(MD \mid MM,t) \\ P(MM \mid MD,t) & P(MD \mid MD,t) \end{pmatrix} \end{array}, \quad (86)$$

which in [35] is parameterized with two real numbers $r \geq 0$, and $0 \leq a \leq 1$ as

$$\begin{array}{c} \\ MM \\ MD \end{array} \begin{array}{c} \begin{array}{cc} MM & MD \end{array} \\ \begin{pmatrix} a + (1-a)e^{-rt} & (1-a)(1-e^{-rt}) \\ a - ae^{-rt} & 1 - a + ae^{-rt} \end{pmatrix} \end{array}. \quad (87)$$

Tree-HMMs model the evolution of paths though the HMM. In contrast, the method proposed here models the evolution of the transition probabilities of the model themselves. However, one can see that the match-transition matrix is closely related in form to the model we have proposed here. Introduce the probability vectors,

$$q_t^{MM} = \begin{pmatrix} P(MM \mid MM,t) \\ P(MD \mid MM,t) \end{pmatrix}, \; q_t^{MD} = \begin{pmatrix} P(MM \mid MD,t) \\ P(MD \mid MD,t) \end{pmatrix}. \quad (88)$$

For $t = 0$ and $t = \infty$ they have the following values,

$$q_0^{MM} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \; q_0^{MD} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \; q_\infty^{MM} = q_\infty^{MD} = \begin{pmatrix} a \\ 1-a \end{pmatrix}. \quad (89)$$

It is easy to see that the match-transition matrix given in equation (87) can be rewritten as,

$$q_t^{MM} = q_\infty^{MM} + e^{tR}(q_0^{MM} - q_\infty^{MM}), \quad (90)$$

$$q_t^{MD} = q_\infty^{MD} + e^{tR}(q_0^{MD} - q_\infty^{MD}), \quad (91)$$

for a diagonal rate matrix $R = \begin{pmatrix} -r & 0 \\ 0 & -r \end{pmatrix}$. Such diagonal rate matrix does not require additional normalization because it corresponds to the case described in equation (80).

Derivation of the algorithm to evolve a vector of transition probabilities

To describe the evolution of transition probabilities, the simple exponential models used for substitution matrices are not sufficient. I propose to adopt a generalization of the exponential model of the form,

$$q_t^T = q_0^T + r^T(e^{tR} - I), \quad (92)$$

where I is the $n \times n$ identity matrix, r is a vector still to be identify, and a R is a $n \times n$ rate matrix.

This model simply adds to the exponential term a time independent vector a,

$$q_t^T = a^T + r^T e^{tR}. \quad (93)$$

Because $q_{t=0} = q_0$, then it is necessary that $a = q_0 - r$, thus giving the expression in equation (92). Note that this is the most general solution of a differential equation of the form $\dot{q}_t \propto (q_t - a)$. Until now it was always assumed that the constant term was zero, that is $r = q_0$. The freedom added by including a term constant in time is that, while before the behavior at $t = \infty$ was solely controlled by e^{tR}, now the additional term also contributes to that limit.

An immediate consequence of this generalization is that the rate matrix is not now sufficient to determine the whole evolutionary process. In addition to the rate matrix, the probability vector must also be specified at time zero (no divergence) and at time infinity (all mutations have saturated) such that,

$$q_\infty^T = q_o^T + r^T\left[\lim_{t\to\infty} e^{tR} - I\right]. \quad (94)$$

The exponential of the rate matrix R has the general form,

$$\begin{aligned} e^{tR} &= \exp\left\{-t\, U \begin{pmatrix} k_1 & & \\ & \ddots & \\ & & k_n \end{pmatrix} U^{-1}\right\} \\ &= U \begin{pmatrix} e^{-k_1 t} & & \\ & \ddots & \\ & & e^{-k_n t} \end{pmatrix} U^{-1}, \end{aligned} \quad (95)$$

for some real eigenvalues $\{k_i\}_{i=1}^n$. If conditions are restricted to the case in which $k_i > 0$, $\forall i$, the immediate consequence of working with positive eigenvalues is that,

$$\lim_{t\to\infty} e^{tR} = 0. \qquad (96)$$

There is then a simple relationship between the vector r and the values of the probabilities at time zero and saturation,

$$q_\infty = q_0 - r. \qquad (97)$$

Therefore, we can write with all generality

$$\begin{aligned} q_t^T &= q_0^T + r^T[e^{tR} - I] \\ &= q_\infty^T + (q_0 - q_\infty)^T e^{tR} \\ &= q_\infty^T + (q_0 - q_\infty)^T U \begin{pmatrix} e^{-k_1 t} & & \\ & \ddots & \\ & & e^{-k_n t} \end{pmatrix} U^{-1}. \end{aligned} \qquad (98)$$

However, for the given information (q_0, q_*, q_∞), the time-parameterized vector q_t in (98) is still underdetermined.

In order to reduce the amount of freedom, I assume that e^{tR} is diagonal (*i.e.* $U = I$). Diagonal rate matrices have been used in other contexts of generalized evolution such as the tree-HMM model [34,35]. Then we have,

$$q_t^T = q_\infty^T + (q_0 - q_\infty)^T \begin{pmatrix} e^{-k_1 t} & & \\ & \ddots & \\ & & e^{-k_n t} \end{pmatrix}. \qquad (99)$$

At this time the known probabilities at the generating time t_*, q_*, have not yet been used. These are,

$$q_*^T = q_\infty^T + (r(1)e^{-k_1 t_*}, \ldots, r(n)e^{-k_n t_*}). \qquad (100)$$

Thus we obtain

$$e^{-k_i t_*} = \frac{q_*(i) - q_\infty(i)}{q_0(i) - q_\infty(i)} = 1 + \frac{q_*(i) - q_0(i)}{q_0(i) - q_\infty(i)}, \qquad (101)$$

which can be solved for k_i,

$$k_i = -\frac{1}{t_*} \log \frac{q_*(i) - q_\infty(i)}{q_0(i) - q_\infty(i)}. \qquad (102)$$

The condition $k_i > 0$ translates into $0 < e^{-k_i t^*} < 1$, or

$$-1 < \frac{q_*(i) - q_0(i)}{q_0(i) - q_\infty(i)} < 0. \qquad (103)$$

This condition has two solutions:

$$q_0(i) < q_*(i) < q_\infty(i) \text{ or } q_0(i) > q_*(i) > q_\infty(i). \qquad (104)$$

Even though this model was derived under the conditions of equation (104), it also extends to the degenerate case where for some i we have

$$q_0(i) = q_*(i) = q_\infty(i), \qquad (105)$$

since this simply corresponds to these parameters undergoing no evolution at all.

Therefore if the input column vectors satisfy one of the three previous conditions for each one of their elements, the parametric expression is

$$\begin{aligned} q_t^T &= q_\infty^T + (q_0 - q_\infty)^T e^{tR} \\ &= q_\infty^T + (q_0 - q_\infty)^T \begin{pmatrix} \exp\left[t/t_* \log\left(\frac{q_*(1) - q_\infty(1)}{q_0(1) - q_\infty(1)}\right)\right] & & \\ & \ddots & \\ & & \exp\left[t/t_* \log\left(\frac{q_*(n) - q_\infty(n)}{q_0(n) - q_\infty(n)}\right)\right] \end{pmatrix}. \end{aligned} \qquad (106)$$

A normalization condition has not yet been imposed. Using the unity vector $u^T = (1, \ldots, 1)$, normalization requires that

$$q_t^T u = 1 \quad \forall t. \qquad (107)$$

For an evolutionary model of the form $q_t \propto e^{tR}$ normalization requires that $e^{tR}u = u$ for arbitrary times. I refer to this property as the strong normalization condition. The normalization of a generalized evolutionary model of the form $q_t^T = q_\infty^T + (q_0 - q_\infty)^T e^{tR}$ requires the weaker condition $(q_0 - q_\infty)^T e^{tR} u = 0$. This property is always true for a rate matrix that satisfies the strong normalization condition. I refer to this property as the weak normalization condition.

In order to obtain the strong normalization condition automatically it is necessary to have a rate matrix of the form $R = U\, diag\,(0, -\lambda_2, \ldots, -\lambda_n)\, U^{-1}$ (see Appendix A, equation (193)). Such a type of rate matrix is not appropriate to describe the evolution of a probability vector, since such rate matrix cannot be uniquely inferred from the three input probability vectors q_0, q_*, q_∞. For that reason, I have explored the use of rate matrices of the diagonal form $R = diag\,(-k_1, \ldots, -k_n)$, which can be inferred from the three input probability vectors q_0, q_*, q_∞ using expression (102). Such diagonal rate matrices, however, in general do not satisfy the strong normalization condition, thus the weak normalization normalization condition must be obtained by other means.

Define w_t:

$$w_t \equiv (q_0 - q_\infty)^T e^{tR} u = \sum_{i=1}^{n} [q_0(i) - q_\infty(i)] \exp\left[t/t_* \log\left(\frac{q_*(i) - q_\infty(i)}{q_0(i) - q_\infty(i)}\right)\right]. \quad (108)$$

If the input vectors satisfy the condition for all i,

$$\frac{q_*(i) - q_\infty(i)}{q_0(i) - q_\infty(i)} = -r, \quad (109)$$

for some real number $r > 0$, then the rate matrix has the particular form $R = diag(-r,..., -r)$, the function $w_t = 0$ for arbitrary times, and the weak normalization condition is satisfied automatically.

The previous condition is in general too restrictive. If the previous condition is not satisfied, by construction w_t is zero at $t = 0$, $t = t_*$ and $t = t_\infty$, but in general $w_t \neq 0$ for $n > 2$. Normalization is then achieved (107) by modifying our definition of q_t to

$$q_t \leftarrow \frac{q_t}{1 + w_t}$$

The final expression is

$$q_t^T = \frac{q_\infty^T}{1+w_t} + \frac{(q_0 - q_\infty)^T}{1+w_t} \begin{pmatrix} \exp[t/t_* \log(\frac{q_*(1)-q_\infty(1)}{q_0(1)-q_\infty(1)})] & & \\ & \ddots & \\ & & \exp[t/t_* \log(\frac{q_*(n)-q_\infty(n)}{q_0(n)-q_\infty(n)})] \end{pmatrix}. \quad (110)$$

Evolution of a matrix of transition probabilities

In some cases, several states of a model correspond to a particular evolutionary event, and it seems natural to expect that their transitions would evolve under the control of the same rate matrix. For instance, in a profile HMM (Figure 4) I will consider the joint evolution of the transitions of three states associated with a given consensus position: Match (M), Insert (I), and Delete (D).

For a collection of m states $S = \{S_1,..., S_m\}$ that transition into a collection of n states $E = \{E_1,..., E_n\}$, consider the set of all transition probabilities emerging from the m originating states S and ending in the n states E,

$$T^{S_{\tilde{i}} \to E_i}, \text{ for } \tilde{i} = 1,...,m, \text{ and } i = 1,...,n. \quad (111)$$

The set of S and E states do not have to be mutually exclusive, and some E states can also be part of the S set. The set of E states also has to be complete, in the sense that

$$\sum_{i=1}^{n} T^{S_{\tilde{i}} \to E_i} = 1. \quad \text{for all } S_{\tilde{i}}. \quad (112)$$

On the other hand, not all E states need to be reached by a given $S_{\tilde{i}}$ state; some transitions may be forbidden by design. For instance, for the states associated with a consensus position in a profile HMM the set of originating states is $S = \{M, D, I\}$, and the set of ending states is $E = \{M', D', I\}$, where the prime index indicates the next position in the profile. The condition $m \leq n$ can be imposed with all generality.

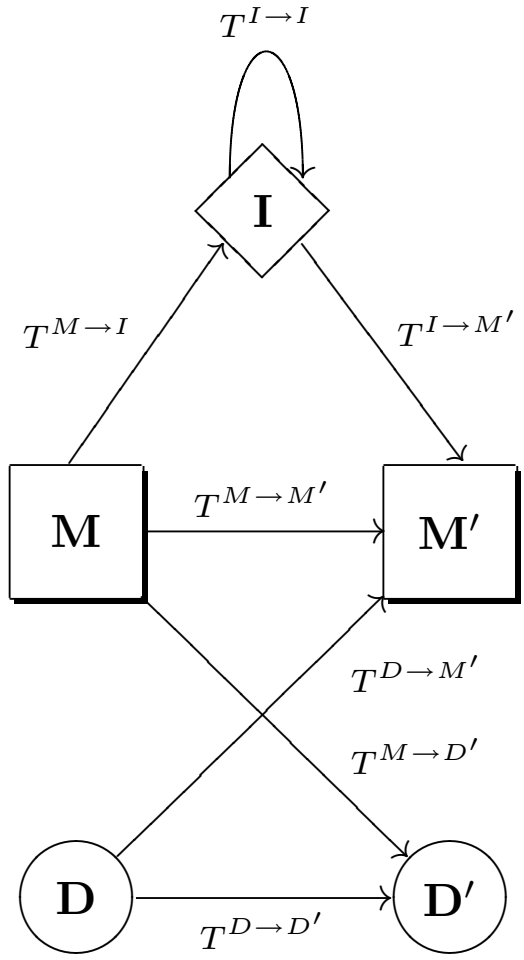

Figure 4
Part of a profile HMM model. For a profile HMM we depict the transition probabilities associated with the states of a given consensus position in the profile: Match (*M*), Insert (*I*), and Delete (*D*). The three states corresponding to the next position in the profile are referred to with primes. The Match state has three transitions (into *I*, *M*', and *D*'), while the Insert and Delete states have two transitions each (into *I* and *M*', and into *D*' and *M*' respectively).

We want to describe the evolution with time of the transition probabilities. For that purpose, I will use the $m \times n$ matrix of transitions Q_t such that,

$$Q_t(\tilde{ii}) \equiv T_t^{S_{\tilde{i}} \to E_i}. \qquad (113)$$

Note that an evolutionary model of the form $Q_t = Q_0 e^{tR}$, like that used for evolving gaps as extra characters, is not sufficient. A model of the form $Q_t = Q_0 e^{tR}$ in the limit of infinite divergence would necessarily result in transitions that for a given end state are all identical and independent of the previous state. That is clearly too restrictive for most models, for instance in a profile HMM, in which some of the transitions are not evolved and are set to zero.

In order to allow for more general saturation properties of the transition probabilities, I propose the following model for the evolution of the matrix of transition probabilities,

$$Q_t = Q_0 + K\,(e^{tR} - I), \qquad (114)$$

where R is the $n \times n$ rate matrix, and the $m \times n$ matrix K is still to be determined. This extension (as in the vector model proposed before) corresponds to adding a constant term $A = Q_0 - K$, and it is is the more general solution of a differential equation of the form $\dot{Q}_t \propto (Q_t - A)$.

We will see that $K(e^{tR} - I) = (Q_0 - Q_\infty)(e^{tR} - I)$, thus

$$Q_t = Q_0 + (Q_0 - Q_\infty)\,(e^{tR} - I_{n \times n}), \qquad (115)$$

$$= Q_\infty + (Q_0 - Q_\infty)\,e^{tR}. \qquad (116)$$

As in the previous case, a freedom provided by the additional constant-in-time term is that while the saturation behavior of $Q_0 e^{tR}$ is controlled by the saturation probabilities of e^{tR}, the model given by equation (116) is independent of those saturation probabilities so that the probabilities at infinity can be set arbitrarily. That is, assuming that ψ is the n dimensional vector of saturation probabilities of e^{tR},

$$\lim_{t\to\infty} Q_0 e^{tR} = Q_0 u_n \psi^T = u_n \psi^T, \qquad (117)$$

$$\lim_{t\to\infty} [Q_\infty + (Q_0 - Q_\infty)e^{tR}] = Q_\infty + (Q_0 - Q_\infty)u_n\psi^T = Q_\infty. \qquad (118)$$

Notice that while Q_t, Q_0, Q_∞ and K are $m \times n$ matrices operating in the $S \times E$ space, the matrices R and e^{tR} are square $n \times n$ matrices operating in the $E \times E$ space. In fact, e^{tR} determines the change in time that a transition probability into one of the E states experiences and in which fashion that change is absorbed by the transition probabilities into any other E state.

A step-by-step description of the algorithm

The recipe to implement the algorithm is as follows:

1. Assume we know the $m \times n$ ($m \le n$) matrices of transition probabilities at time zero Q_0 and at time infinity Q_∞, such that the rank of $Q_0 - Q_\infty$ is m.

2. If an analytic $n \times n$ rate matrix R is given, one can find the analytic expression for e^{tR} by solving the differential equation $d(e^{tR})/dt = Re^{tR}$, and jump to step (6). For a numerical solution jump to step (5).

3. If the information given is the set of transition probabilities at a generating time t_*, calculate the rate matrix R as,

$$R = \frac{1}{t_*}\log[I_{n\times n} + O(Q_* - Q_0)] = \frac{1}{t_*}\sum_{l=1}^{l=\infty}\frac{(-1)^{l+1}}{l}[O(Q_* - Q_0)]^l. \qquad (119)$$

The $n \times m$ matrix O is obtained by solving the set of linear equations

$$-(Q_\infty - Q_0)\,O + u_m v^T = I_{m \times m}, \qquad (120)$$

where u_m is the m dimensional unity vector [*i.e.* $u_m^T = (1,\ldots,1)$], and v is a m dimensional vector uniquely determined by the set of m independent linear equations,

$$v^T (Q_\infty - Q_0) = 0, \qquad (121)$$

$$v^T u_m = 1. \qquad (122)$$

The solution of equation (120) is not unique. In fact, equation (120) determines the matrix O up to a n dimensional probability vector ψ that satisfies the conditions $\psi^T O = 0$. This probability vector corresponds to the saturation probabilities of the matrix e^{tR}. While the rate matrix R and the matrix e^{tR} depend on the choice of the saturation probabilities ψ, the asymptotic behaviour of the matrix of transition probabilities is independent of ψ, as was shown in equation (118).

4. Impose the condition,

$$v^T (Q_* - Q_0) = 0. \qquad (123)$$

This condition [necessary so that $\psi^T R = 0$] imposes constraints between the set of probabilities at time zero, at time t_*, and at time infinity.

5. Calculate the exponential of the rate matrix e^{tR} using the corresponding Taylor expansion.

6. Finally, calculate the set of evolved transition probabilities as,

$$Q_t = Q_0 - (Q_\infty - Q_0)(e^{tR} - I_{n \times n}) \quad (124)$$

or

$$Q_t = Q_\infty + (Q_0 - Q_\infty)\, e^{tR}. \quad (125)$$

An example: Evolution of the transitions of a profile HMM given a rate matrix

To illustrate this method, consider the case of a profile HMM (Figure 4). There are three states associated with a given consensus position in the profile: the Match state (M), the Insert state (I) and the Delete state (D). These three states transition into states M' (the Match state at the next position in the profile), D' (the delete state at the next position in the profile), and I (the insert state between the two matched positions), therefore in this example $m = 3$ and $n = 3$.

Consider the following the transition matrix

$$Q_t = \begin{array}{c} \\ M \\ D \\ I \end{array} \begin{array}{c} \begin{array}{ccc} M' & D' & I \end{array} \\ \begin{pmatrix} T_t^{M\to M'} & T_t^{M\to D'} & T_t^{M\to t} \\ T_t^{D\to M'} & T_t^{D\to D'} & 0 \\ T_t^{I\to M'} & 0 & T_t^{I\to I} \end{pmatrix} \end{array}. \quad (126)$$

This transition matrix, like a nucleotide substitution matrix, adds up to one by rows. We assume as in HMMER [4] that there are no transitions between the Insert and the Delete states, but the model could work under more general conditions.

The matrix at time zero is given by

$$Q_t = \begin{array}{c} \\ M \\ D \\ I \end{array} \begin{array}{c} \begin{array}{ccc} M' & D' & I \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 1-q_D & q_D & 0 \\ 1-q_I & 0 & q_I \end{pmatrix} \end{array}. \quad (127)$$

The parameter $1/(1 - q_I)$ is the average length of an insert in between two matched positions at very short times (if there were no deletions). The parameter $1/(1 - q_D)$ is the average length of a deletion at very short times (if there were no insertions, and all position in the profile had the same parameters at time zero). For instance, one could set q_I very close to zero, which implies that, for very small times when $T_{\delta t}^{M\to I} > 0$, the average length of a insertion would be very close to one.

At time infinity, one can parameterize the transition probabilities as

$$Q_\infty = \begin{array}{c} \\ M \\ D \\ I \end{array} \begin{array}{c} \begin{array}{ccc} M' & D' & I \end{array} \\ \begin{pmatrix} 1-(m_D+m_I) & m_D & m_I \\ 1-d_D & d_D & 0 \\ 1-i_I & 0 & i_I \end{pmatrix} \end{array}. \quad (128)$$

where m_D and m_D represent the probabilities of Match to Delete and Match to Insert at infinity, and d_D and i_I are the Delete to Delete and Insert to Insert probabilities at time infinity, $(0 \le m_D, m_I, d_D, i_I \le 1)$.

Let us assume that the rate matrix is given by

$$R = \begin{array}{c} \\ M' \\ D' \\ I \end{array} \begin{array}{c} \begin{array}{ccc} M' & D' & I \end{array} \\ \begin{pmatrix} -2\alpha & \alpha & \alpha \\ \alpha & -2\alpha & \alpha \\ \alpha & \alpha & -2\alpha \end{pmatrix} \end{array}. \quad (129)$$

for some parameter $\alpha > 0$. This rate matrix assumes that the rate of change in the occurrence of state M' is similar to that of state D' and that of state I, and that this change reverts equally into the other two states. More realistic situations can be achieved using rate matrices depending on more parameters.

The instantaneous rate of transition change is given by,

$$-(Q_\infty - Q_0)R = \begin{array}{c} \\ M \\ D \\ I \end{array} \begin{array}{c} \begin{array}{ccc} M' & D' & I \end{array} \\ \begin{pmatrix} -3\alpha(m_D+m_I) & 3\alpha m_D & 3\alpha m_I \\ -3\alpha(d_D-q_D) & 3\alpha(d_D-q_D) & 0 \\ -3\alpha(i_I-q_I) & 0 & 3\alpha(i_I-q_I) \end{pmatrix} \end{array}. \quad (130)$$

This matrix gives the instantaneous change that a transition probability experiences under this model and describes how that change is transferred to the other allowed transition probabilities.

The matrix e^{tR} can be obtained analytically by solving the differential equation $d(e^{tR})/dt = Re^{tR}$. This is a 3-dimensional Jukes-Cantor model which has as its solution,

$$e^{tR} = \begin{array}{c} \\ M' \\ D' \\ I \end{array} \begin{array}{c} \begin{array}{ccc} M' & D' & I \end{array} \\ \begin{pmatrix} r_t & s_t & s_t \\ s_t & r_t & s_t \\ s_t & s_t & r_t \end{pmatrix} \end{array}. \quad (131)$$

with

$$r_t = \frac{1}{3} + \frac{2}{3}e^{-3\alpha t}, \qquad (132)$$

$$s_t = \frac{1}{3} - \frac{1}{3}e^{-3\alpha t}. \qquad (133)$$

Putting all together, we obtain the following evolved transition probabilities for a profile HMM under a Jukes-Cantor like assumption for the rate matrix:

$$Q_t = Q_\infty + (Q_0 - Q_\infty)e^{tR} = \begin{array}{c} \\ M \\ D \\ I \end{array} \begin{array}{c} \begin{array}{ccc} M' & D' & I \end{array} \\ \begin{pmatrix} 1-3(m_D+m_I)s_t & 3m_D s_t & 3m_I s_t \\ (1-q_D)-3(d_D-q_D)s_t & q_D+3(d_D-q_D)s_t & 0 \\ (1-q_I)-3(i_I-q_I)s_t & 0 & q_I+3(i_I-q_I)s_t \end{pmatrix} \end{array}. \qquad (134)$$

Substituting the values for s_t given in equation (133), we have the following evolved transition probabilities for a profile HMM,

$$T_t^{M\to M'} = 1-(m_I+m_D)(1-e^{-3\alpha t}), \qquad (135)$$

$$T_t^{M\to D'} = m_D(1-e^{-3\alpha t}), \qquad (136)$$

$$T_t^{M\to I} = m_I(1-e^{-3\alpha t}), \qquad (137)$$

$$T_t^{D\to M'} = (1-q_D)-(d_D-q_D)(1-e^{-3\alpha t}), \qquad (138)$$

$$T_t^{D\to D'} = q_D+(d_D-q_D)(1-e^{-3\alpha t}), \qquad (139)$$

$$T_t^{I\to M'} = (1-q_I)-(i_I-q_I)(1-e^{-3\alpha t}), \qquad (140)$$

$$T_t^{I\to I} = q_I+(i_I-q_I)(1-e^{-3\alpha t}), \qquad (141)$$

The evolution of different paths through the HMM

In a tree-HMM one assumes that the different paths through the model are the objects that are subject to evolution [34]. Here we have directly modeled the evolution of the transition probabilities of the HMM. We can get an intuition for the meaning of these evolved transition probabilities by estimating how these evolved transition probabilities induce the evolution of different paths through the model. A process that is similarly to that modeled by a one-branch tree-HMM.

Suppose that at time zero, we emitted residue *a* from state *M*, and residue *b* from state *M*'. The model assigns to such sequence a probability given by,

$$P(ab \mid t=0) = p_M(a)p_{M'}(b)T_0^{M\to M'} = p_M(a)p_{M'}(b), \qquad (142)$$

where $p_M(a)$ and $p_{M'}(b)$ represent the emission probabilities associated to the *M* and *M*' states respectively. Now suppose that at time *t* there has been an insertion of *n* residues in between the two matches *a* and *b*; the model assigns to such sequence a probability given by,

$$P(ai_1\ldots i_n b \mid t) = p_M(a)p_{M'}(b)p_I(i_1)\ldots p_I(i_n)T_t^{M\to I}(T_t^{I\to I})^{n-1}T_t^{I\to M'}, \qquad (143)$$

where $p_I(i_i)$ represent the emission probabilities associated to the Insert state.

We can interpret that in time *t* the path through the model that generated *ab* has evolved into the path through the model that generated $ai_1\ldots i_n b$ with probability given by

$$P(ai_1\ldots i_n b \mid ab,t) = p_I(i_1)\ldots p_I(i_n)T_t^{M\to I}(T_t^{I\to I})^{n-1}T_t^{I\to M'}. \qquad (144)$$

To get a better intuition of what this means, take as an example the case in which the time interval *t* is very small.

Then the probability that a path between two matches in the HMM inserts *n* residues in time $t \approx 0$ is

$$P(ai_i\ldots i_n b \mid ab,t) \approx p_I(i_1)\ldots p_I(i_n)3\alpha m_I q_I^{n-1}(1-q_I)t. \qquad (145)$$

This probability is proportional to $3\alpha m_I$ the rate of substituting a Match-to-Match transition for a Match-to-Insert transition, and to q_I^{n-1} (1 - q_I), which is the geometric factor associated to an insert of length *n* at time zero.

An example: Evolution of the transitions of a profile HMM given the transitions at a generating time

In this case, we maintain the same values for the transition probabilities at time zero Q_0 and at time infinity Q_∞, but the rate matrix will be obtained from a generating time for which we know the transition probabilities.

The set of linear equations in step (3) of this algorithm that determine the vector $v^T = (v_1, v_2, v_3)$ are

$$m_D v_1 + (d_D - q_D)v_2 = 0, \qquad (146)$$

$$m_I v_1 + (i_I - q_I)v_3 = 0, \qquad (147)$$

$$v_1 + v_2 + v_3 = 1. \qquad (148)$$

The solution of these linear equations is

$$v_1 = (d_D - q_D)(i_I - q_I)/d, \qquad (149)$$

$$v_2 = -m_D(i_I - q_I)/d, \qquad (150)$$

$$v_3 = -m_I(d_D - q_D)/d, \qquad (151)$$

where $d \equiv (d_D - q_D)(i_I - q_I) - m_D(i_I - q_I) - m_I(d_D - q_D)$.

Parameterize the matrix *O* in the form

$$O = \begin{pmatrix} M_1 & M_2 & M_3 \\ D_1 & D_2 & D_3 \\ I_1 & I_2 & I_3 \end{pmatrix}, \qquad (152)$$

with each row adding to zero. The set of linear equations in step (3) that determine the matrix O are

$$(d_D - q_D)(M_1 - D_1) + v_1 = 0,\ (i_I - q_I)(M_1 - I_1) + v_1 = 0, \qquad (153)$$

$$(d_D - q_D)(M_2 - D_2) + v_2 = 1,\ (i_I - q_I)(M_2 - I_2) + v_2 = 0, \qquad (154)$$

$$(d_D - q_D)(M_3 - D_3) + v_3 = 0,\ (i_I - q_I)(M_3 - I_3) + v_3 = 1. \qquad (155)$$

Solving M_i and I_i in terms of D_i we have,

$$O = \begin{pmatrix} D_1 - \frac{1}{d}(i_I - q_I) & D_2 + \frac{1}{d}(i_I - q_I - m_I) & D_3 + \frac{m_I}{d} \\ D_1 & D_2 & D_3 \\ D_1 + \frac{1}{d}[(d_D - q_D) - (i_I - q_I)] & D_2 + \frac{1}{d}(i_I - q_I - m_I - m_D) & D_3 - \frac{1}{d}(d_D - q_D - m_D - m_I) \end{pmatrix}. \qquad (156)$$

The matrix O is therefore determined up to the unitary vector (D_1, D_2, D_3). The saturation probabilities $\psi^T = (\psi_{M'}, \psi_{D'}, \psi_{I'})$ $(\psi^T R = 0)$ are defined by the equations $\psi^T O = 0$, which imply

$$D_1 = \psi_M, \frac{i_I - q_I}{d} - \psi_I \frac{(d_D - q_D) - (i_I - q_I)}{d}, \qquad (157)$$

$$D_2 = -\psi_{M'} \frac{i_I - q_I - m_I}{d} - \psi_I \frac{i_I - q_I - m_I - m_D}{d}, \qquad (158)$$

$$D_3 = -\psi_{M'} \frac{m_I}{d} + \psi_I \frac{d_D - q_D - m_D - m_I}{d}. \qquad (159)$$

Substituting vector D with vector ψ we finally obtain the following expression for the matrix O in terms of the saturation probabilities ψ:

$$O = \frac{1}{d}\begin{pmatrix} -\psi_{D'}(i_I - q_I) - \psi_I(d_D - q_D) & \psi_{D'}(i_I - q_I) + \psi_I m_D - \psi_{D'} m_I & \psi_I(d_D - q_D) + \psi_{D'} m_I - \psi_I m_D \\ (\psi_{M'} + \psi_I)(i_I - q_I) - \psi_I(d_D - q_D) & -(\psi_{M'} + \psi_I)(i_I - q_I - m_I) + \psi_I m_D & \psi_I(d_D - q_D - m_D) - (\psi_{M'} + \psi_I) m_I \\ (\psi_{M'} + \psi_{D'})(d_D - q_D) - \psi_{D'}(i_I - q_I) & \psi_{D'}(i_I - q_I - m_I) - (\psi_{M'} + \psi_{D'}) m_D & -(\psi_{M'} + \psi_{D'})(d_D - q_D - m_D) + \psi_{D'} m_I \end{pmatrix}. \qquad (160)$$

The condition in step (4) of the algorithm translates in this case into the following relationship of parameters,

$$(d_D - q_D) t_*^{M \to D'} = m_D (t_*^{D \to D'} - q_D), \qquad (161)$$
$$(i_I - q_I) t_*^{M \to I} = m_I (t_*^{I \to I} - q_I).$$

This is an additional set of constraints that the "vector" algorithm does not impose.

To test the algorithm I have made up a toy HMM consensus state, which at the generating time t_* is given by the matrix of transitions,

$$Q_* = \begin{pmatrix} 0.70 & 0.20 & 0.10 \\ 0.60 & 0.40 & 0 \\ 0.70 & 0 & 0.30 \end{pmatrix}. \qquad (162)$$

Selecting the particular values $q_D = q_I = 0.1$ and $d_D = i_I = 0.6$, using the constraints of equations (161) implies that $m_D = 0.33$ and $m_I = 0.25$. Using these values and the arbitrary values for the saturation probabilities $\psi = (1/3, 1/3, 1/3)$, we obtain the following O matrix:

$$O = \begin{pmatrix} 8.0000 & -4.6667 & -3.3333 \\ -4.0000 & 1.3333 & 2.6667 \\ -4.0000 & 3.3333 & 0.6667 \end{pmatrix}. \qquad (163)$$

The rate matrix R constructed using equation (119) is given by

$$R = \begin{pmatrix} -0.4757 & 0.3054 & 0.1703 \\ 0.4406 & -0.6109 & 0.1703 \\ 0.0351 & 0.3054 & -0.3406 \end{pmatrix}, \qquad (164)$$

and an instantaneous rate matrix $-(Q_\infty - Q_0)R$ is given by,

$$-(Q_\infty - Q_0)R = \begin{pmatrix} -0.4323 & 0.3046 & 0.1277 \\ -0.4569 & 0.4569 & 0 \\ -0.2554 & 0 & 0.2554 \end{pmatrix}. \qquad (165)$$

The evolved set transition probabilities at time $t = 0.3$ is given by,

$$Q_{0.3} = \begin{pmatrix} 0.8845 & 0.0800 & 0.0355 \\ 0.7800 & 0.2200 & 0 \\ 0.8290 & 0 & 0.1710 \end{pmatrix}. \qquad (166)$$

Using the "vector" method, in which transition probability vectors evolve independently with the same set of parameters, we would have obtained the identical result,

$$Q_{0.3} = \begin{pmatrix} 0.8846 & 0.0800 & 0.0354 \\ 0.7798 & 0.2202 & 0 \\ 0.8290 & 0 & 0.1710 \end{pmatrix}. \qquad (167)$$

The normalization function w_t given in equation (82) is different from zero only for dimensions larger than two. The second and third row effectively have dimension two (since one of the elements is always zero), and do not require normalization. For the first row the normalization function takes the value $w_{0.3} = 0.0020$.

The vector method allows us to use more unrestricted sets of parameters than the matrix method since the conditions in equation (78) are independent for each row. In principle, however, the conditions in equation (161)

seem to allow behaviors that the vector model does not allow such as $t_*^{M\to D'} > m_D \equiv t_\infty^{M\to D'}$ as long as, simultaneously, $t_*^{D\to D'} > d_D \equiv t_\infty^{D\to D'}$. In practice when I have tested that kind of situation, the rate matrices obtained are always not real, and therefore they lack any biological interpretation.

Derivation of the algorithm to evolve a matrix of transition probabilities

We start with a model of the general form

$$Q_t = Q_0 + K\,(e^{tR} - I_{n\times n}), \quad (168)$$

where Q_0 is the known $m \times n$ matrix of probabilities at time zero, and the $m \times n$ matrix K must still be determined.

Assume that we know the transition probabilities at time infinity, which we represent by the $m \times n$ matrix Q_∞. Then, because of the asymptotic behavior of the exponential family e^{tR}, $\lim_{t\to\infty} e^{tR} = u_n \psi^T$ for some n dimensional saturation probabilities, where $\psi^T = (\psi_i, \ldots, \psi_n)$, and the n dimensional unity vector $u_n^T = (1, \ldots, 1)$, we have

$$Q_\infty = Q_0 + K\,(u_n \psi^T - I_{n\times n}). \quad (169)$$

This equation implies that

$$K = -(Q_\infty - Q_0) + \tilde{k}\,\psi^T, \quad (170)$$

where $\tilde{k}$ is a m dimensional vector that represents the sums by rows of K, *i.e.* $\sum_j K(i, j) = \tilde{k}_i$ which we impose to be different from zero.

Because for the exponential family e^{tR} we have the reversibility condition $\psi^T e^{tR} = \psi^T$ for arbitrary time, introducing the expression for K in equation (170) in the equation (114) we have the general result,

$$Q_t = Q_0 - (Q_\infty - Q_0)(e^{tR} - I_{n\times n}). \quad (171)$$

This result proves point (6) of the previous algorithm description.

Therefore if given Q_0, Q_∞ and a $n \times n$ rate matrix R, which satisfy the reversibility conditions $\psi^T R = 0$, we can calculate the evolved transition probabilities using the equation (171).

In the case in which the information given is the set of transition probabilities at a generating time t_*, designated by Q_*, the calculation of the rate matrix R involves the following steps:

(a) The $m \times n$ matrix K is by construction invertible because we have imposed $\tilde{k}_i \neq 0$, for all rows i.

A little aside with respect to matrix inversions is in order here. The (unique) inverse of a matrix is defined only for square matrices. One can introduce a inverse-like matrix for a non-square matrix; these are called pseudoinverses [69]. The pseudoinverse of a non-square matrix is not unique and many pseudoinverses can be defined; one of the best known is the Moore-Penrose matrix inverse [70]. We will see how despite the fact that the pseudoinverse of K is not unique, we can still define Q_t uniquely.

Therefore solving for R in equation (114) at the particular time t_* we have

$$R = \frac{1}{t_*}\log\left[\, I + K^{-1}\left(Q_* - Q_0\right)\right], \quad (172)$$

where K^{-1} is the $n \times m$ pseudoinverse of K defined by the conditions $KK^{-1} = I_{m\times m}$ and $K^{-1}K = I_{n\times n}$.

(b) Because the final result for Q_t in equation (116) does not depend on the values $\tilde{k}_i$ we can set them with all generality to the form $\tilde{k}_i = \rho \neq 0$. Therefore we have

$$K = -(Q_\infty - Q_0) + \rho u_m\, \psi^T. \quad (173)$$

Because $K^{-1}Ku_n = u_n$ and $Ku_n = \rho u_m$, then we need that $K^{-1}u_m = \rho^{-1}u_n$. Therefore we propose that the $n \times m$ pseudoinverse matrix K^{-1} has the following form,

$$K^{-1} = O + \frac{1}{\rho} u_n v^T, \quad (174)$$

where the $n \times m$ matrix O, and the m dimensional vector v satisfy the conditions,

$$O\, u_m = 0, \quad (175)$$

$$v^T u_m = 1. \quad (176)$$

(c) In order to satisfy $K^{-1}K = \mathrm{I}_{n\times n}$ we need to have,

$$v^T(Q_\infty - Q_0) = 0, \quad (177)$$

$$-O(Q_\infty - Q_0) + u_n \psi^T = I_{n\times n}. \quad (178)$$

Equation (177) is a set of homogeneous linear equations that together with the normalization conditions in equation (176) uniquely determine the vector v.

On the other hand, in order to satisfy $KK^{-1} = I_{m\times m}$, the following must apply:

$$\psi^T O = 0, \quad (179)$$

$$-(Q_\infty - Q_0)O + u_m v^T = I_{m\times m}. \quad (180)$$

Equation (180) is a set of linear equations which determines O aside from a dependence on an arbitrary probability vector. In particular we can find the expression of matrix O in terms of the vector ψ as we did in equation (160).

Once the matrix O has been obtained using equation (180) as a function of the vector ψ, one can verify that the set of equations describe by (178) is automatically satisfied for any vector ψ as long as it satisfies the condition $\psi^T O = 0$. This is the result presented in step (4) of the algorithm.

(d) Because ψ corresponds to the saturation probabilities of e^{tR}, then it is necessary that $\psi^T R = 0$. This condition is satisfied if,

$$\psi^T [O + \frac{1}{\rho} u_n v^T](Q_* - Q_0) = 0. \quad (181)$$

Therefore it implies that,

$$v^T(Q_* - Q_0) = 0, \quad (182)$$

which is the condition imposed in step (4) of the algorithm. Under those conditions, it results for the rate matrix R,

$$R = \frac{1}{t_*}\log\left[I_{n\times n} + O(Q_* - Q_0)\right]. \quad (183)$$

Notice that the parameter $\rho \neq 0$, which is necessary to be able to invert the matrix K to calculate the rate R, does not appear anywhere in the final result, either in the evolved transitions Q_t or in the value of R. This results from the fact that in either equation the only relevant component is the projection of K (or K^{-1}) into $(e^{tR} - I)$. The same projection is what makes the vector ψ that appears in the pseudoinverse K^{-1} irrelevant. Even though $\lim_{t\to\infty} e^{tR} = u\psi^T$, it is also true that $\lim_{t\to\infty} (Q_\infty - Q_0)e^{tR} = 0$, so that the dependence on ψ disappears from the final expression of Q_∞.

Reversibility and multiplicativity

For a given probabilistic model, imposing reversibility has different implications for its emission and transition probabilities. In pair models, we assume that the emission probabilities are reversible by imposing $P(a_t, b_{t+t'}) = P(a_{t+t'}, b_t)$, which corresponds to using symmetric joint probabilities represented by the shorthand notation $P(a, b|t')$. If the emissions do not involve gaps, the marginal probabilities do not evolve, and the evolved joint probabilities are obtained from the evolved conditionals and the saturation probabilities. In the presence of gaps, I have described how to construct the evolved conditionals and the corresponding evolved marginals in a way that maintains reversibility for any arbitrary time, so that we can construct evolved symmetric joint probabilities.

For transition probabilities the situation is different. Mathematically, a matrix of transition probabilities is like a substitution matrix (*i.e.* conditional probabilities) but there is not the equivalent of "joint" probabilities for transitions. To maintain reversibility for the transitions of a probabilistic model, one has to build reversibility in the design of the model. In particular, one needs to be sure that the transition probabilities that involve gaps lack any directionality. For instance, in the pair-HMM of Figure 3 we need to impose that $T_t^{XY\to X} = T_t^{XY\to Y}$ for arbitrary times. That is achieved by making sure that the input transition probabilities at time t_*, zero and infinity do lack directionality.

Another property of probabilistic models of evolution for residue substitutions is multiplicativity. Multiplicativity is an immediate property for evolutionary models of the form e^{tR}. For residue-substitution evolutionary processes, multiplicativity implies that the transition from one given event (say residue a) to another event (say residue b) in a finite time, if it goes through any intermediate state, has to be of the form of any other possible substitution. In mathematical terms,

$$P(b|a, t + t') = \sum_c P(c|a, t)P(b|c, t'). \quad (184)$$

However, when allowing gaps, any intermediate evolutionary step can go through processes of deletions or insertions in addition to substitutions; therefore multiplicativity as described in the previous equation does not hold anymore. There is a natural explanation of why "substitutions-only multiplicativity" is modified when considering insertion and deletion events. Consider the evolution of gaps as single characters, which was introduced previously in this paper. The substitution matrix with gaps Q_t^{ε} satisfies the relationship

$$Q^{\varepsilon}_{t+t'} = Q^{\varepsilon}_{t} Q_0^{-1} Q^{\varepsilon}_{t'}. \qquad (185)$$

Analyzing this matrix equation by components and using the expression for Q_0 given in equation (22), the substitution of residue a into residue b in finite time $t + t'$ has the following terms:

$$\begin{aligned} P(b|a,t+t') = &\sum_c P(c|a,t) P(b|c,t') \\ &+ \frac{1}{q_0} P(-|a,t) P(b|-,t') \\ &+ \left(1 - \frac{1}{q_0}\right) P(-|a,t) \sum_c p_c P(b|c,t'). \end{aligned} \qquad (186)$$

The first term corresponds to pure substitution events of the form $a \xrightarrow{t} c \xrightarrow{t'} b$, and it is identical to equation (184). The second term modulated by the coefficient $1/q_0$ (introduced in equation (65), which is part of the non trivial matrix Q_0) represents the event in which $a \xrightarrow{t} - \xrightarrow{t'} - b$. The third term (preceded by coefficient (1 - $1/q_0$)) represents the event in which $a \xrightarrow{t} - c \xrightarrow{t'} - b$. Note that this model would align at time $t + t'$ residues which could have been derived by a gap intermediate. This is usually discouraged by evolutionary models that describe the evolutionary history of insertions and deletions, in which such event would be represented as $\begin{matrix} a & - \\ - & b \end{matrix}$. For the model at hand, the fact that a gap can revert into a residue is a consequence of treating gaps as an additional residue in a substitution matrix.

For the particular case of the generalized Jukes-Cantor model introduced before, it turns out that the two extra terms in equation (186) are independent of the particular substitutions and cancel, such that

$$\frac{1}{q_0} \gamma_t \xi_{t'} + \left(1 - \frac{1}{q_0}\right) \gamma_t \frac{1 - \gamma_{t'}}{4} = 0. \qquad (187)$$

Therefore the generalized Jukes-Cantor model preserves multiplicativity. This results from the extreme simplicity of the model and is not true for more complicated models. For instance, for the rate matrix created from a particular Q_* in the other example presented in this paper (which is a particular case of the REV model [44]), the two extra terms in equation (186) are different for the different nucleotide substitutions, and do not cancel out.

A more complicated situation appears for probabilistic models that introduce gaps in an affine manner. A given residue-to-residue substitution process that occurred in a finite time could have appeared from a very large number of intermediate situations in which stretches of other nucleotides could have been added or removed. The simple one-to-one correspondence that models of substitutions maintain through evolution does not exist in the presence of insertion and deletion events. This does not mean that evolutionary models with gaps are inconsistent, however some traditional properties of phylogenetic trees of single residue evolution such as the pulley principle [71] cannot be applied under the transition probability evolution models.

Conclusion

Motivated by the goal of making QRNA (a comparative probabilistic method for RNA genefinding) an evolutionary model, I have introduced several probabilistic methods to describe the evolution of insertion and deletion events. The methods introduced here have a larger scope than this program alone, and they can be applied to other pair probabilistic models and to profile HMMs and SCFGs as well.

I described an algorithm which addresses the evolution of gaps as an extra residue in a $(N + 1) \times (N + 1)$ substitution matrix. This method can be applied to the joint emission probabilities of pair models. This method allows us to maintain a stationary N-dimensional background distribution, while the actual $(N + 1)$-dimensional background frequencies evolve towards all gaps at time infinity. I call this process quasi-stationary. As an example, I showed an analytic solution for the Jukes-Cantor model extended to gaps.

I also presented two methods for the evolution of transition probabilities in a profile or pair HMM or SCFG, that are applicable to any probabilistic model that uses transitions between states to model insertions and deletions. In the first algorithm, the transition probabilities associated with one state in the model are evolved as a vector independently of the transition probabilities associated to any other state in the model. I also presented a second algorithm in which the transition probabilities associated with a given set of states co-evolve under the control of a single rate matrix. I presented an example of the application of these methods to a pair-HMM and to a profile HMM.

I have applied these methods to the program QRNA, which was the motivation for the development of the algorithms in the first place. QRNA contains three probabilistic models (the oth, cod, and rna models) that analyze the pattern of mutation of a given pairwise alignment to decide which of the three models best classifies the alignment. These models are a combination of generalized pair-HMMs and a pair-SCFG. Originally, this program assumed a fixed divergence time, and all the

emission probabilities of the different models were tied to those of BLOSUM62. That produced a QRNA parameterized for highly diverse sequences, which in turn produced a large number of false positives for highly similar sequences. In the new program eQRNA, all emission and transition probabilities are a continuous time-dependent family able to match any possible degree of sequence divergence.

The three models of QRNA (the OTH, COD, and RNA models) need to be at approximately the same evolutionary distance, so that when a pairwise alignment is analyzed, the differences in scores of the models result from observing a different pattern of mutations (coding, RNA, or none in particular) rather than because one model favors more closely related sequences than the other. This model synchronization requires a number of QRNA-specific design elements which are tangential to the implementation of the evolutionary models for indels and transition probabilities presented in this paper. For reasons of clarity, I leave for another paper a detailed description of the particular implementation designs that went into eQRNA in order to make it fully evolutionary. In a nutshell, the transition probabilities of the OTH and COD models are evolved according to the algorithm to evolve vectors of transition probabilities, while the emission probabilities of those two models were evolved using the original QRNA parameters as the generating time of the respective rate matrix. In the RNA model, for the context-free grammar component of the model, the transitions are fixed, and the evolution of gaps is accommodated by treating gaps as extra characters according to the method presented here for that purpose. The HMM component of the RNA model is parameterized with time similarly to the OTH and COD models. Preliminary results show an important improvement compared with the previous fixed-time implementation. The application of these evolutionary methods for other probabilistic models for sequence comparison beyond eQRNA should be tractable.

So far the methods presented here have been introduced only in profile and pair models. They could also be applied to probabilistic models where, instead of aligning two contemporary sequences, one aligns a sequence to an ancestor. The only difference with respect to an evolutionary pair model is that, in this case, the emission probabilities will be the substitution (conditional) matrices themselves instead of joint conditional-on-time probabilities. One important limitation of the methods presented here is that, in general, they lack the property of multiplicativity. In consequence, in order to extend the methods presented here to more than two sequences related by a phylogenetic tree, one would have to work with rooted trees. A future challenge is to incorporate these evolutionary methods into multiple sequence probabilistic models that explicitly describe the phylogenetic relationship between the sequences.

Availability

The different models presented in this paper have been implemented in several small ANSI C programs. These are not fully developed software applications, but demonstrations (for those who want to avoid the mathematical descriptions) of how the different algorithms work. The programs are freely available at http://selab.wustl.edu/publications/Rivas05/evolve.tar.gz.

Methods

Appendix A. Conditions for the saturation of a generalized substitution matrix

In this appendix I provide the conditions for saturation of a generalized evolutionary model of the form $Q_t = Q_0 e^{tR}$. Saturation can be described as

$$\lim_{t\to\infty} Q_t = \begin{pmatrix} q_\infty^1 & \cdots & q_\infty^n \\ \vdots & \vdots & \vdots \\ q_\infty^1 & \cdots & q_\infty^n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \left(q_\infty^1, \ldots, q_\infty^n \right) = u q_\infty^T, \qquad (188)$$

for the unitary vector u, and a set of saturation frequencies at time infinity denoted by q_∞, such that $q_\infty^T u = 1$.

Here I show that saturation of $Q_t = Q_0 e^{tR}$ is a necessary condition of two properties of the matrix $Q = \{Q(ij)\}$, normalization and positivity. I also show that the saturation probabilities of Q_t are the same as those of e^{tR}.

Proposition A.1. Consider first the simplest case $Q_t = e^{tR}$. Normalization, *i.e.* $\sum_j Q(ij) = 1$, together with positivity, *i.e.* $Q(ij) > 0\ \forall i, j$, imply that a substitution matrix of the form $Q_t = e^{tR}$ saturates to a set of probabilities at time infinity.

Proof. Normalization of the rate matrix, $\sum_j Q(ij) = 1$ implies that

$$Qu = Q \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}. \qquad (189)$$

That is, $\lambda = 1$ is an eigenvalue of Q. It also has implications for the norm of Q, defined as the largest row sum of absolute values

$$\|Q\| \equiv \max_i \sum_{j=1}^{n} |Q(ij)| = 1. \qquad (190)$$

Therefore, because of the spectral theorem [72], the spectral radius $\sigma(Q)$, defined as the largest absolute value of any eigenvalue of Q, is bounded by,

$$\sigma(Q) \leq ||Q|| = 1. \quad (191)$$

On the other hand there is an eigenvalue $\lambda = 1$ therefore

$$\sigma(Q) = 1. \quad (192)$$

In consequence, Q has one eigenvalue, $\lambda = 1$, and all other eigenvalues are smaller than one.

Therefore because the substitution matrix is of the form $Q_t = e^{tR}$, it implies that the instantaneous rate matrix R has one null eigenvalue, and all the other are negative. If we assume that the null eigenvalue is not degenerate and that the negative eigenvalues are real, we can write with all generality,

$$R = U \begin{pmatrix} 0 & & & \\ & -\lambda_2 & & \\ & & \ddots & \\ & & & -\lambda_n \end{pmatrix} U^{-1}, \quad (193)$$

for some matrix U, and such that $\lambda_i > 0$ for $i = 2, \ldots, n$.

Therefore $Q_t = e^{tR}$ can be cast into the form,

$$Q_t = U \begin{pmatrix} 1 & & & \\ & e^{-t\lambda_2} & & \\ & & \ddots & \\ & & & e^{-t\lambda_n} \end{pmatrix} U^{-1}. \quad (194)$$

In the limit,

$$\lim_{t\to\infty} Q_t = U \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} (1,0,\ldots,0) U^{-1} = U\Psi_0\Psi_0^T U^{-1}, \quad (195)$$

for $\Psi_0^T = (1, 0, \ldots, 0)$.

On the other hand using equation (194) we obtain

$$Q_t U\Psi_0 = U \begin{pmatrix} 1 & & & \\ & e^{-t\lambda_2} & & \\ & & \ddots & \\ & & & e^{-t\lambda_n} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = U \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \ \forall t, \quad (196)$$

which implies that $U\Psi_0$ is the eigenvector of Q corresponding to the eigenvalue $\lambda = 1$. According to (189) that is,

$$U\Psi_0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}. \quad (197)$$

Substituting in equation (195) we finally obtain,

$$\lim_{t\to\infty} Q_t = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \Psi_0^T U^{-1}. \quad (198)$$

This is the saturation condition (188) for some saturation probabilities defined by $q_\infty = \Psi_0^T U^{-1}$.

Corollary A.1. For a generalized evolutionary model of the form $Q_t = Q_0 e^{tR}$, Q_t also saturates at infinity, and the saturation probabilities of Q_t are given by those of e^{tR}, that is,

$$\lim_{t\to\infty} Q_t = \lim_{t\to\infty} e^{tR} = uq_\infty^T. \quad (199)$$

Proof. Note that by construction Q_0 has to have the same normalization and positivity conditions as Q_t. It can be shown that under those conditions, $Q_0^{-1} Q_t = e^{tR}$ also has to add up to one, summing by rows, and all its elements have to be positive. Therefore, using the result of Proposition A.1,

$$\lim_{t\to\infty} Q_0^{-1} Q_t = uq_\infty^T. \quad (200)$$

Therefore

$$\lim_{t\to\infty} Q_t = (Q_0 u) q_\infty^T = uq_\infty^T, \quad (201)$$

which proves saturation for an evolutionary probabilistic process of the form $Q_t = Q_0\, e^{tR}$.

Appendix B. Implications of reversibility on a generalized evolutionary process

In this appendix I discuss the implications that reversibility imposes on a generalized evolutionary model. I show that for an evolutionary model of the form $Q_t = e^{tR}$, the marginal probabilities with respect to which Q_t is reversible have to be stationary, and therefore coincide with the saturation probabilities. I also show that for an evolutionary model of the general form $Q_t = Q_0 e^{tR}$, the marginal probabilities with respect to which Q_t is reversible can change with time. In this way we decouple the "reversibility" frequencies from the saturation frequencies. I also

demonstrate how to calculate the saturation probabilities, given Q_0 and Q_* at one particular time t_*. This system sets the ground for the quasi-stationary model of evolution with gaps as an extra indel.

Lemma B.1. Consider a given matrix of conditional probabilities Q_*, $[\sum_j Q_*(ij) = 1 \ \forall i]$ which is reversible with respect to a set of marginal probabilities p_*,

$$p_*(i)Q_*(ij) = p_*(j)Q_*(ji). \quad (202)$$

Then one can see that reversibility implies

$$p_*^T Q_* = p_*^T. \quad (203)$$

Proof. Summing one of the indices in the reversibility conditions and taking into account the normalization condition for the Q_* matrix results in,

$$\sum_i p_*(i)Q_*(ij) = p_*(j)\sum_i Q_*(ji) = p_*(j), \quad (204)$$

which in vectorial notation takes the form $p_*^T Q_* = p_*^T$. (203)

Lemma B.2. If $R = \log Q_*$ then the reversibility condition (202) for Q_* implies that

$$p_*(i)R(ij) = p_*(j)R(ji) \quad \wedge \quad p_*^T R = 0. \quad (205)$$

Proof. If $R = \log Q_*$ then because of the Taylor series we have

$$R = \frac{1}{t_*}\sum_{n=1}^{n=\infty} \frac{(-)^{n+1}}{n} (Q_* - I)^n. \quad (206)$$

Because of the reversibility condition for Q_* (202) it is also true that

$$p_*(i)\ (Q_* - I)^n\ (ij) = p_*(j)(Q_* - I)^n\ (ji), \quad (207)$$

for $n \geq 1$. Therefore it follows that

$$p_*(i)\ R(ij) = p_*(j)R(ji). \quad (208)$$

In addition we can also see by inspecting equation (206) that the normalization condition for Q_* translates into $\sum_j R(ij) = 0\ \forall i$, which implies that

$$\sum_i p_*(i)R(ij) = 0 \quad \forall j \quad \text{or} \quad p_*^T R = 0. \quad (209)$$

Lemma B.3. If Q_* is a conditional matrix that satisfies the reversibility condition (202) and $R = \log Q_*$ then the saturation probabilities of R are given by the p_* vector in (202), that is,

$$\lim_{t\to\infty} e^{tR} = up_*^T. \quad (210)$$

Proof. Taking from *Lemma B.2.*, we have $p_*^T R = 0$; therefore $p_*^T R^n = 0$ for $n \geq 0$, and because of the relationship

$$e^{tR} = I + \sum_{n=1}^{n=\infty} \frac{(tR)^n}{n!}, \quad (211)$$

it results that

$$p_*^T e^{tR} = p_*^T. \quad (212)$$

for arbitrary t. Therefore, it also holds in the limit of very large time that

$$p_*^T \lim_{t\to\infty} e^{tR} = p_*^T. \quad (213)$$

Additionally, Appendix A shows that $\lim_{t\to\infty} e^{tR} = uq_{sat}^T$. Combining those two equations together we have

$$p_*^T = p_*^T \lim_{t\to\infty} e^{tR} = p_*^T uq_{sat}^T = q_{sat}^T. \quad (214)$$

This proves that the saturation probabilities are p_*.

Proposition B.1. For a reversible evolutionary model of the form $Q_t = e^{tR}$, it results that the associated marginal probabilities with respect to which the parametric family Q_t is reversible have to be stationary (*i.e.* time independent).

Proof. From the parametric family Q_t select one particular instance t_*, and consider $Q_* = Q_{t=t_*}$. Suppose that the marginal probabilities at this time are given by p_*, that is: $p_*^T Q_* = p_*^T$. Because of the relationship $R = \log Q_*$, it follows from Lemma B.3 that the whole parametric family e^{tR} has p_* as the corresponding marginal probabilities, therefore the marginal probabilities do not evolve with time (stationary).

Proposition B.2. For a reversible generalized evolutionary model of the form $Q_t = Q_0 e^{tR}$, the associated marginal probabilities with respect of which Q_t is reversible can be evolved with time.

proof. In order to prove that this is the case, we just need to find an example in which that statement is true. Consider again one particular instance $Q_* = Q_{t=t_*}$ with its corresponding marginal probabilities p_*. Because the model is reversible for arbitrary divergence times, in particular there should be some p_0 probabilities such that $p_0^T Q_0 = p_0^T$. For this generalized model, the rate matrix is given by $R = \log(Q_0^{-1}Q_*)$. Therefore it follow by Lemma B.3 that the saturation probabilities of R are given by the condition

$$p_\infty(i)(Q_0^{-1}Q_*)(ij) = p_\infty(j)(Q_0^{-1}Q_*)(ji). \quad (215)$$

Therefore the saturation probabilities p_∞ are different from p_* as long as $p_0 \neq p_*$.

Therefore, we have constructed a parametric family, $Q_t = Q_0 e^{tR}$, in which the marginal probabilities for reversibility are p_0 at time zero, p_* at t_*, and p_∞ at time infinity, with $p_0 \neq p_* \neq p_\infty$. Therefore if there is reversibility at arbitrary time, the marginals have to be time dependent,

$$p_t(i)Q_t(ij) = p_t(j)Q_t(ji). \quad (216)$$

In particular in the Section "The evolution of emission probabilities with indels treated as an extra character" we have constructed a system in which the time-dependent reversibility condition (216) is satisfied by marginal probabilities that are quasi-stationary with respect to some $(n - 1)$ p_0 probabilities,

$$p_t(i) = p_0(i)\ (1 - \Lambda_t), \text{ for } i = 1,...n - 1 \quad (217)$$

$$p_t(n) = \Lambda_t. \quad (218)$$

Acknowledgements

Thanks to Sean Eddy for numerous discussions. Thanks to Matt Visser for insights into matrix logarithms. This work was supported by the NIH National Human Genome Research Institute. I would like to acknowledge the Benasque Center for Science in which part of this work took shape at an ESF and NIH funded workshop on computational RNA biology in the summer of 2003.

References

1. Durbin R, Eddy SR, Krogh A, Mitchison GJ: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* Cambridge UK: Cambridge University Press; 1998.
2. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology: Applications to protein modeling.** *J Mol Biol* 1994, **235:**1501-1531.
3. Karplus K, Barrett C, Hughey R: **Hidden Markov models for detecting remote protein homologies.** *Bioinformatics* 1998, **14:**846-856.
4. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
5. Eddy SR: **Multiple Alignment Using Hidden Markov Models.** In *Proc Third Int Conf Intelligent Systems for Molecular Biology* Edited by: Rawlings C, Clark D, Altman R, Hunter L, Lengauer T, Wodak S. Menlo Park, CA: AAAI Press; 1995:114-120.
6. Burge CB, Karlin S: **Finding the Genes in Genomic DNA.** *COSB* 1998, **8:**346-354.
7. Cawley SL, Pachter L: **HMM sampling and applications to gene finding and alternative splicing.** *Bioinformatics* 2003:II36-II41. ii36–ii41
8. Meyer IM, Durbin R: **Gene structure conservation aids similarity based gene prediction.** *Nucl Acids Res* 2004, **32:**776-783.
9. Sakakibara Y, Brown M, Hughey R, Mian IS, Sjolander K, Underwood RC, Haussler D: **Stochastic Context-Free Grammars for tRNA Modeling.** *Nucl Acids Res* 1994, **22:**5112-5120.
10. Eddy SR, Durbin R: **RNA Sequence Analysis Using Covariance Models.** *Nucl Acids Res* 1994, **22:**2079-2088.
11. Lowe TM, Eddy SE: **tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence.** *Nucl Acids Res* 1997, **25:**955-964.
12. Eddy SR: **A Memory-Efficient Dynamic Programming Algorithm for Optimal Alignment of a Sequence to an RNA Secondary Structure.** *BMC Bioinformatics* 2002, **3:**18.
13. Klein RJ, Eddy SR: **RSEARCH: Finding homologs of single structured RNA sequences.** *BMC Bioinformatics* 2003, **4:**44.
14. Knudsen B, Hein J: **RNA secondary structure prediction using stochastic context-free grammars and evolutionary history.** *Bioinformatics* 1999, **15:**446-454.
15. Dowell RD, Eddy SR: **Evaluation of Several Lightweight Stochastic Context-Free Grammars for RNA Secondary Structure Prediction.** *BMC Bioinformatics* 2004, **5:**71.
16. Rivas E, Eddy SR: **Noncoding RNA gene detection using comparative sequence analysis.** *BMC Bioinformatics* 2001, **2:**8.
17. Altschul S, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool.** *Jour Mol Biol* 1990, **215:**403-410.
18. Yang Z: **Estimating the pattern of nucleotide substitution.** *J Mol Evol* 1994, **39:**105-111.
19. Goldman N, Thorne JL, Jones DT: **Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses.** *J Mol Biol* 1996, **263:**196-208.
20. Muse SV: **Estimating synonymous and nonsynonymous substitution rates.** *Mol Biol Evol* 1996, **13:**105-114.
21. Whelan S, Goldman N: **general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach.** *Mol Biol Evol* 2001, **18:**691-699.
22. Smith AD, Lui TW, Tillier ER: **Empirical models for substitution in ribosomal RNA.** *Mol Biol Evol* 2004, **21:**419-421.
23. Knudsen B, Andersen ES, Damgaard C, Kjems J, Gorodkin J: **Evolutionary rate variation and RNA secondary structure prediction.** *Comput Biol Chem* 2004, **28:**219-226.
24. Yang Z: **A space-time process model for the evolution of DNA sequences.** *Genetics* 1995, **139:**993-1005.
25. Felsenstein J, Churchill GA: **Hidden Markov Model approach to variation among sites in rate of evolution.** *Mol Biol Evol* 1996, **13:**93-104.
26. Gribskov M, Veretnik S: **Identification of sequence pattern with profile analysis.** *Methods Enzymol* 1996, **266:**198-212.
27. Coin L, Durbin R: **Improved techniques for the identification of pseudogenes.** *Bioinformatics* 2004:I94-I100.
28. McAuliffe JD, Pachter L, Jordan MI: **Multiple-sequence functional annotation and the generalized hidden Markov phylogeny.** *Bioinformatics* 2004, **20:**1850-1860.
29. Siepel A, Haussler D: **Combining phylogenetic and hidden Markov models in biosequence analysis.** *J Comput Biol* 2004, **11:**413-428.
30. Thorne JL, Kishino H, Felsenstein J: **An evolutionary model for maximum likelihood alignment of DNA sequences.** *J Mol Evol* 1991, **33:**114-124.
31. Thorne JL, Kishino H, Felsenstein J: **Inching toward reality: an improved likelihood model of sequence evolution.** *J Mol Evol* 1992, **34:**3-16.
32. Metzler D: **Statistical alignment based on fragment insertion and deletion models.** *Bioinformatics* 2003, **19:**490-499.
33. Miklos I, Lunter GA, Holmes I: **"Long Indel" model for evolutionary sequence alignment.** *Mol Biol Evol* 2004, **21:**529-540.
34. Mitchison GJ, Durbin RM: **Tree-based maximal likelihood substitutions matrices and hidden Markov models.** *J Mol Evol* 1995, **41:**1139-11351.

35. Mitchison GJ: **probabilistic treatment of phylogeny and sequence alignment.** *J Mol Evol* 1999, **49:**11-22.
36. Holmes I, Bruno W: **Evolutionary HMMs: a bayesian approach to multiple alignment.** *Bioinformatics* 2001, **17:**803-820.
37. Qian B, Goldstein RA: **Detecting distant homologs using phylogenetic tree-based HMMs.** *Proteins* 2003, **52:**446-453.
38. Holmes I: **Using guide trees to construct multiple-sequence evolutionary HMMs.** *Bioinformatics* 2003, **Suppl 1:**147-157.
39. Knudsen B, Miyamoto MM: **Sequence alignments and pair hidden Markov models using evolutionary history.** *J Mol Biol* 2003, **333:**453-460.
40. Pedersen JS, Hein J: **Gene finding with a hidden Markov model of genome structure and evolution.** *Bioinformatics* 2003, **19:**219-227.
41. Holmes I: **A probabilistic model for the evolution of RNA structure.** *BMC Bioinformatics* 2004, **5:**166.
42. Jukes TH, Cantor C: **Evolution of protein molecules.** In *Mamm Prot Met* Academic Press; 1965:21-132.
43. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16:**111-120.
44. Tavaré S: **Some probabilistic and statistical problems in the analysis of DNA sequences.** *Lectures on Mathematics in the Life Sciences* 1986, **17:**57-86.
45. Yang Z, Nielsen R, Hasegawa M: **Models of amino acid substitution and applications to mitochondrial protein evolution.** *Mol Biol Evol* 1998, **15:**1600-1611.
46. Kosiol C, Goldman N, Buttimore NH: **new criterion and method for amino acid classification.** *J Theor Biol* 2004, **228:**97-106.
47. Yang Z, Nielsen R, Goldman N, Pedersen A: **Codon-substitution models for heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155:**431-449.
48. Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985, **21:**160-174.
49. Holmes I, Rubin GM: **An expectation maximization algorithm for training hidden substition models.** *J Mol Biol* 2002, **317:**757-768.
50. Müller T, Spang R, Vingron M: **Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood methods.** *Mol Biol Evol* 2002, **19:**8-13.
51. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci* 1992, **89:**10915-10919.
52. Kishino H, Miyata T, Hasegawa M: **Maximum likelihood inference of protein phylogeny and the origin of chloroplasts.** *J Mol Evol* 1990, **31:**151-160.
53. Dayhoff M, Schwartz R, Orcutt B: **model of evolutionary change in protein.** *Atlas Prot Seq Struct* 1978, **5:**345-352.
54. Müller T, Vingron M: **Modeling amino acid replacement.** *J Comp Biol* 2000, **7:**761-776.
55. Kosiol C, Goldman N: **Different Versions of the Dayhoff Rate Matrix.** *Mol Biol Evol* 2004, **22:**193-199.
56. Israel RB, Rosenthal JS, Wei JZ: **Finding generators for Markov chains via empirical transition matrices, with applications to credit rating.** *Mathematical Finance* 2001, **11:**245-265.
57. Kreinin A, Sidelnikova M: **Regularization algorithms for transition matrices.** *Algo Res Quartely* 2001, **4:**23-40.
58. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author.** *Department of Genome Sciences, University of Washington, Seattle* 2004.
59. Swofford DL: **PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.** *Sinauer Associates, Sunderland, Massachusetts* 2003.
60. Adachi J, Hasegawa M: **MOLPHY programs for molecular phylogenetics version 2.3.** *Institute of Statistical Mathematics, Tokyo* 1995.
61. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13:**555-556.
62. Liò P, Goldman N, Thorne JL, Jones3 DT: **PASSML: combining evolutionary inference and protein secondary structure prediction.** *Bioinformatics* 1998, **14:**726-733.
63. Ronquist F, Huelsenbeck JP: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754-755.
64. Cai W, Pei J, Grishin NV: **Reconstruction of ancestral protein sequences and its applications.** *BMC Evol Biol* 2004, **4:**33.
65. Siepel A, Haussler D: **Phylogenetic estimation of context-dependent substitution rates by maximum likelihood.** *Mol Biol Evol* 2004, **21:**468-488.
66. Lunter G, Hein J: **A nucleotide substitution model with nearest-neighbour interactions.** *Bioinformatics* 2004:I216-I223.
67. Goldman N, Whelan S: **A novel use of equilibrium frequencies in models of sequence evolution.** *Mol Biol Evol* 2002, **19:**1821-1831.
68. Whelan S, Goldman N: **Estimating the frequency of events that cause multiple-nucleotide changes.** *Genetics* 2004, **167:**2027-2043.
69. Campbell SL, Meyer CDJ: *Generalized Inverses of Linear Transformations* New York: Dover; 1991.
70. Jodár L, Law AG, Rezazadeh A, Watson JH, Wu G: **Computations for the Moore-Penrose and Other Generalized Inverses.** *Congress Numer* 1991, **80:**57-64.
71. Felsenstein J: **Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach.** *J Mol Evol* 1981, **17:**368-376.
72. Bronson R: *Matrix operations* New York: McGraw-Hill; 1973.