BIOINFORMATICS Bioinformatics Advance Access published February 24, 2005

Using evolutionary Expectation Maximisation to estimate indel rates

Ian Holmes [a,b]

[a]*Present address: Department of Bioengineering, University of California, Berkeley CA 94720-1762, USA,* [b]*Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK*

ABSTRACT

Motivation: The Expectation Maximisation algorithm, in the form of the Baum-Welch algorithm (for HMMs) or the Inside-Outside algorithm (for SCFGs), is a powerful way to estimate the parameters of stochastic grammars for biological sequence analysis. To use this algorithm for multiple-sequence evolutionary modeling, it would be useful to apply the EM algorithm to estimate not just the probability parameters of the stochastic grammar, but also the instantaneous mutation rates of the underlying evolutionary model (to facilitate the development of stochastic grammars based on phylogenetic trees, also known as *Statistical Alignment*). Recently, we showed how to do this for the point substitution component of the evolutionary process; here, we extend these results to the indel process.

Results: We present an algorithm for maximum likelihood estimation of insertion and deletion rates from multiple sequence alignments, using Expectation Maximisation, under the single-residue indel model due to Thorne, Kishino and Felsenstein (the "TKF91" model). The algorithm converges extremely rapidly, gives accurate results on simulated data that are an improvement over parsimonious estimates (which are shown to underestimate the true indel rate), and gives plausible results on experimental data (coronavirus envelope domains). Due to the algorithm's close similarity to the Baum-Welch algorithm for training Hidden Markov Models, it can be used in an "unsupervised" fashion to estimate rates for unaligned sequences, or estimate several sets of rates for sequences with heterogenous rates.

Availability: Software implementing the algorithm and the benchmark is available under GPL from http://www.biowiki.org/

Contact: Ian Holmes ihh@berkeley.edu

Keywords: Statistical alignment, Evolutionary HMMs, TKF model

1 INTRODUCTION

A cornerstone of stochastic grammar-based profiling of DNA, protein and RNA sequences is the fast and accurate estimation of the probability parameters of the model, given a "training set". This is achieved for hidden Markov models (HMMs) using the Baum-Welch algorithm, and for stochastic context-free grammars (SCFGs) using the Inside-Outside algorithm [1]; both are examples of the Expectation Maximisation algorithm [2].

Currently, most HMM and SCFG-based methods are consensus- or pair-based; they do not use the underlying phylogenetic tree directly in their training (although consensus methods use it indirectly, when over-represented clades are downweighted in the estimation of the consensus profile [1]). The full power of evolutionary methods to model mutation rates and phylogenetic correlations is therefore unused. This may be because evolutionary models relating multiple sequences, using an arbitary phylogeny, are more complicated than consensus models, which effectively assume a star-shaped phylogeny (with the consensus sequence at the centre of the star). Nonetheless, there has been some progress in implementing stochastic evolutionary models for multiple sequence alignment and profiling [3, 4, 5, 6, 7, 8, 9]. To develop this work further, a natural step is to extend the probabilistic results used for HMMs and SCFGs to the evolutionary domain.

Previous work demonstrated the potential of the EM algorithm as a quick and effective tool for finding maximum-likelihood rate matrices for point substitution processes [7]. As with all EM algorithms, this algorithm proceeds by repeated application of an E-step followed by an M-step. During the E-step, limited information about the inferred evolutionary history is gathered in the form of certain "sufficient statistics", which are then used during the M-step to improve the estimate of the rate parameters. Specifically, these sufficient statistics are (i) the expected composition of the root sequence $\hat{s}_i$, (ii) the expected number of times $\hat{u}_{ij}$ that each mutation of type $i \to j$ occurred and (iii) the expected length of time $\hat{w}_i$ that each residue was available for substitution. All expectations are taken over the posterior distribution of substitution histories. In the M-step, equilibrium frequencies are set proportional to $\hat{s}_i$ and substitution rates equal to $\hat{u}_{ij}/\hat{w}_i$. This process converges extremely rapidly and, assuming the

absence of significantly suboptimal local minima in the likelihood function (which in practise is often the case), results in maximum-likelihood substitution rate matrices.

This paper presents an EM algorithm for the estimation of indel rates in the TKF91 model of sequence evolution [10]. The main insight is that the analogues of the sufficient statistics (i) $\hat{s}_i$ (ii) $\hat{u}_{ij}$ and (iii) $\hat{w}_i$ for the TKF91 indel process are (i) the expected initial sequence length, (ii) the expected number of insertions and deletions and (iii) the expected mean sequence length over the evolutionary interval (together with the length of the interval itself, T). We have implemented the algorithm in freely available software. We show that it accurately and quickly recovers the true indel length on simulated data; in contrast, the "naive" or "parsimonious" estimate that associates every observed gap with a single indel event is shown to underestimate the true indel rate by a small amount. Previous maximum likelihood (ML) parameterisation approaches have relied on sampling methods, which are significantly slower, albeit less prone to local minima in the likelihood function [11]. We report on an application of the algorithm to estimating indel rates in coronavirus proteins.

Using methods such as this rate-based EM algorithm, a number of novel sequence analysis approaches become feasible. For example, it is theoretically possible to do multiple sequence alignment without manually setting scoring parameters; instead, unbiased estimates of these parameters can be obtained quickly, directly and accurately from data. The algorithms described here can be coupled with a recent "structural EM" algorithm for the estimation of the maximum likelihood phylogenetic topology [12]. Another application is phylogenetic profiling: previous workers such as Thorne *et al* have shown that it is possible to partition heterogenous multiple alignments using the varying signature of the substitution process, e.g. to predict secondary structure in proteins [13] or transcription factor binding sites in promoters [14]. Using methods based on the TKF model is now possible to use similar methods to partition the alignment according to the indel rates, and to estimate the relevant heterogenous rate parameters in an unsupervised fashion, from aligned or unaligned sequences, using EM, exactly as one might do with a Single-sequence or Pair HMM [1]. Thus, this algorithm represents a further step in combining the achievments of HMM-based profiling and evolutionary modeling for multiple-sequence comparative genomics.

2 MODELS

We assume some familiarity with continuous-time Markov models for sequence evolution. For a more thorough introduction, the reader is referred to [10] or [15].

Following Thorne *et al* [10], we make the simplifying assumption that the evolutionary model can be separated into independent components: the point substitution process, modeled as a continuous-time finite-state Markov chain at each site of the sequence, and the indel process, modeled as a linear birth-death process with immigration.

The two models share certain features, in that they are both continuous-time Markov models. We begin by introducing some notation for such models, then describe each in detail.

2.1 Continuous-time Markov models

A continuous-time Markov chain is specified by parameters $\vartheta = \{\pi, \mathbf{R}\}$. Here, the initial distribution is π and the transition rate matrix is $\mathbf{R}$.

The probability of the model being in state j at time t, given that it started in state i at time 0, is $M_{ij}(t)$, where $\mathbf{M}(t) = e^{\mathbf{R}t}$ is the *matrix exponential*.

2.2 The point substitution model

The substitution model is a finite-state continuous-time Markov chain. If the substitution model is reversible, then the rate matrix $\mathbf{R}$ obeys detailed balance ($\pi_i R_{ij} = \pi_j R_{ji}$) and is related to a symmetric matrix $\mathbf{S}$ (by $S_{ij} = R_{ij}\sqrt{\pi_i/\pi_j}$). The eigenvalues $\mu^{(k)}$ and orthonormal eigenvectors $\mathbf{v}^{(k)}$ of $\mathbf{S}$ (satisfying $\mathbf{S}\mathbf{v}^{(k)} = \mu^{(k)}\mathbf{v}^{(k)}$ and $\mathbf{v}^{(k)} \cdot \mathbf{v}^{(l)} = \delta_{kl}$) can easily be found using standard algorithms [16], yielding the following result for the entries of $\mathbf{M}(t)$

$$M_{ij}(t) = \left(\frac{\pi_j}{\pi_i}\right)^{\frac{1}{2}} \sum_{k=1}^{m} v_i^{(k)} e^{\mu^{(k)} t} v_j^{(k)} \qquad (1)$$

2.3 The TKF91 indel model

In contrast to the substitution model, the state space for the TKF91 indel model is infinite (the sequence length is unbounded). However, much of the theory from continuous-time Markov models still applies.

The TKF91 model describes the evolution of a single sequence under the action of two kinds of mutation event: (i) point substitutions, which act on a single residue only (this process uses the finite-state Markov chain framework described in the previous section); and (ii) single-residue indels, which insert or delete a single residue [10]. Insertions occur at rate λ per available site; deletions at rate μ.

Consider first the simple example where we ignore the alignment and just look at the sequence length. We then have a *linear birth-death process with immigration* whose state space is the non-negative integers [17]. The rate matrix is sparse: $R_{ij} = (i+1)\lambda$ if $j = i+1$, $R_{ij} = i\mu$ if $j = i-1$ and $R_{ij} = 0$ if $|i-j| > 1$. Thus $R_{ii} = -(i+1)\lambda - i\mu$. The equilibrium distribution over sequence lengths is geometric, $\pi_i = g^i(1-g)$, where $g = \lambda/\mu$. We assume that the process is always started at equilibrium.

To find the likelihood of an individual alignment, consider starting with some ancestral sequence of length L and allowing it to evolve for some time t. A feature of the TKF91 model is that this process can be simplified by splitting the ancestral sequence into a series of independently evolving "links". For a sequence of length L, there are $L+1$ such

links, including one *immortal link* (representing the insertion site at the leftmost end of the sequence) and L *mortal links* (representing each ancestral residue of the sequence, and the insertion site immediately to its right).

Starting from the initial (ancestral) state, we allow each link to evolve stochastically for some time t, and then examine the new (descendant) state of the link. We may describe this descendant state as follows: first, by the number of newly-inserted residues (n); second (for mortal links) by specifying whether the original ancestral residue survived, or was deleted. Let the probability distribution of n be $r_n(t)$ for immortal links and $p_n(t) + q_n(t)$ for mortal links, where p_n accounts for the survival of the ancestral residue and q_n accounts for its deletion. Then it can be shown [10] that

$$\begin{array}{rcll} p_n(t) & = & \alpha(t)\beta(t)^{n-1}\left(1-\beta(t)\right) & \\ q_n(t) & = & \left(1-\alpha(t)\right)\left(1-\gamma(t)\right) & \text{for } n=0 \\ & = & \left(1-\alpha(t)\right)\gamma(t)\beta(t)^{n-1}\left(1-\beta(t)\right) & \text{for } n>0 \\ r_n(t) & = & \beta(t)^n\left(1-\beta(t)\right) & \end{array} \tag{2}$$

where

$$\begin{array}{rcl} \alpha(t) & = & e^{-\mu t} \\ \beta(t) & = & \frac{\lambda(1-e^{(\lambda-\mu)t})}{\mu-\lambda e^{(\lambda-\mu)t}} \\ \gamma(t) & = & 1-\frac{\mu(1-e^{(\lambda-\mu)t})}{(1-e^{-\mu t})(\mu-\lambda e^{(\lambda-\mu)t})} \end{array} \tag{3}$$

This probabilistic model for pairwise alignments can be expressed as a Pairwise Hidden Markov Model using e.g. the notation of Durbin *et al* [1], as shown in Figure 1.

The restriction to single-residue indel events leads to the geometric term β^n in the above expressions, and is therefore roughly equivalent to using a linear gap penalty to score a sequence alignment. This is widely acknowledged to be unrealistic [18, 4, 19, 9]. Models that incorporate more realistic length distributions over indel sequences are less tractable than TKF91: such models have been analysed using simulation and combinatoric approximations, but have not yet yielded any algebraic expression for the alignment likelihood [19, 9]. A tractable alternative is provided by the TKF92 model, which essentially replaces the single residues of TKF91 with artificial, indivisible, multi-residue "fragments". The lengths of these fragments are geometrically distributed. Where TKF91 is a birth-death process on residues, TKF92 is a birth-death process on fragments.

3 ALGORITHMS

We now describe the EM algorithms for estimating model parameters. Again, we start with general theory for continuous-time Markov models, and proceed to the specific cases (the point substitution model and the TKF91 indel model).

3.1 The EM algorithm for continuous-time Markov models

Let $\mathbf{h}$ represent a "history" of the process: that is, a complete specification of the state of the system at all times. The situation we wish to address is the one where we observe some data $\mathbf{O}$ which partially constrains the allowable histories $\mathbf{h}$. For example, we might know what state the process is in at times $t = 0$ and $t = T$, but not in between. The state of the process at times $0 < t < T$ thus constitutes *missing information*.

The EM algorithm [2] consists of maximising the following sum over possible histories with respect to ϑ'

$$\mathcal{Q}(\vartheta,\vartheta') = \sum_{\mathbf{h}} P(\mathbf{h}|\mathbf{O},T,\vartheta)\log P(\mathbf{h},\mathbf{O}|T,\vartheta') \tag{4}$$

Previous work [7] showed that, if we break the time interval $[0,T]$ into small intervals Δt, we obtain

$$\begin{array}{rcl} \mathcal{Q}(\vartheta,\vartheta') & = & \sum_i^m \hat{s}_i \log\pi'_i + \sum_i^m \hat{w}_i R'_{ii} + \sum_i^m \sum_{j\neq i}^m \hat{u}_{ij}\log R'_{ij} \\ & & + f(\vartheta,\Delta t) \end{array} \tag{5}$$

where $\hat{s}_i$ is the expected number of paths that *start* in state i, $\hat{w}_i$ is the expected *wait* in state i (i.e. the amount of time spent in i) and $\hat{u}_{ij}$ is the expected *usage* of transition $i \to j$. Note that $\sum_i \hat{s}_i = 1$ and $\sum_i \hat{w}_i = T$.

Here $f(\vartheta,\Delta t)$ is a function that doesn't depend on the "new" parameters ϑ'. In fact $f(\vartheta,\Delta t) \sim -\log(\Delta t)$.

Ultimately we are interested in the infinitesimal limit $\Delta t \to dt$. In this limit $\hat{s}$, $\hat{w}$ and $\hat{u}$ retain their interpretations as the expected initial state occupancy, the expected waiting time and the expected transition usage. The function f is problematic: $\lim_{\Delta t\to 0} f(\vartheta,\Delta t) = -\infty$. However, f can effectively be dropped from $\mathcal{Q}$, as it disappears when we differentiate w.r.t. ϑ' to maximise the expected log-likelihood of the new parameters.

We want to maximise $\mathcal{Q}$ while ensuring that $\mathbf{R}$ is a valid rate matrix (i.e. $\sum_j R_{ij} = 0$) and π is a normalised probability vector (i.e. $\sum_i \pi_i = 1$).

Introducing these constraints via Lagrange multipliers $\{\mathcal{A}_1 \ldots \mathcal{A}_m, \mathcal{B}\}$ and dropping f, the function to be maximised is

$$\begin{array}{rcl} \mathcal{L}(\vartheta,\vartheta',\alpha,\beta) & = & \sum_i^m \hat{s}_i\log\pi'_i + \sum_i^m \hat{w}_i R'_{ii} \\ & & + \sum_i^m\sum_{j\neq i}^m \hat{u}_{ij}\log R'_{ij} + \sum_i^m \mathcal{A}_i \sum_j^m R'_{ij} \\ & & + \mathcal{B}\left(\sum_i^m \pi'_i - 1\right) \end{array} \tag{6}$$

3.2 The EM algorithm for the substitution model

Suppose that the observed data **O** comprises the initial state a and the final state b. Then

$$\begin{aligned}\hat{s}_i(a,b,T) &= \delta_{ia} \\ \hat{w}_i(a,b,T) &= \frac{\mathcal{I}_{ii}^{ab}(T)}{M_{ab}(T)} \\ \hat{u}_{ij}(a,b,T) &= \frac{\mathcal{I}_{ij}^{ab}(T)}{M_{ab}(T)} R_{ij} \end{aligned} \quad (7)$$

where

$$\begin{aligned}\mathcal{I}_{ij}^{ab}(T) &= \int_{t=0}^{T} M_{ai}(t) M_{jb}(T-t) dt \\ &= \left(\frac{\pi_i \pi_b}{\pi_a \pi_j}\right)^{\frac{1}{2}} \sum_k^m v_a^{(k)} v_i^{(k)} \sum_l^m v_j^{(l)} v_b^{(l)} \mathcal{J}_{kl}(T)\end{aligned}$$

with

$$\mathcal{J}_{kl}(T) = \begin{cases} T e^{(\mu^{(k)} T)} & \text{if } \mu^{(k)} = \mu^{(l)} \\ \frac{e^{(\mu^{(k)} T)} - e^{(\mu^{(l)} T)}}{\mu^{(k)} - \mu^{(l)}} & \text{if } \mu^{(k)} \neq \mu^{(l)} \end{cases}$$

The maximum of $\mathcal{L}$ is at $\pi_i' = \hat{s}_i / \sum_j \hat{s}_j$, $R_{ij}' = \hat{u}_{ij}/\hat{w}_i$ (for $j \neq i$) and $R_{ii}' = -\sum_j R_{ij}'$.

3.3 The EM algorithm for the TKF91 model

We substitute into equation (6) the definitions of $\mathbf{R}$ in terms of λ', μ', given in Section 2.3.

For the equilibrium distribution π, use independent geometric length parameter g, so that $\pi_i = g^i(1-g)$ (as long as the process stays reversible, we will end up with $g = \lambda/\mu$)

$$\begin{aligned}\mathcal{L}(\vartheta, \vartheta') &= \sum_{i=0}^{\infty} \hat{s}_i \left(i \log g + \log(1-g)\right) \\ &\quad - \sum_{i=0}^{\infty} \hat{w}_i \left((i+1)\lambda' + i\mu'\right) \\ &\quad + \sum_{i=1}^{\infty} \hat{u}_{i,i-1} \log(i\mu') \\ &\quad + \sum_{i=0}^{\infty} \hat{u}_{i,i+1} \log((i+1)\lambda') \\ &= \hat{S} \log g + \log(1-g) - (\hat{Z} + T)\lambda' - \hat{Z}\mu' \\ &\quad + \hat{D} \log \mu' + \hat{I} \log \lambda' + f'(\vartheta)\end{aligned}$$

where $\hat{S} = \sum_i \hat{s}_i i$ is the expected initial sequence length, $\hat{Z} = \sum_i \hat{w}_i i$ is the expected total time accrued by mortal links, T is the expected time accrued by the immortal link, $\hat{D} = \sum_i \hat{u}_{i,i-1}$ is the expected number of deletions, $\hat{I} = \sum_i \hat{u}_{i,i+1}$ is the expected number of insertions and $f'(\vartheta)$ does not depend on ϑ'.

Note the Lagrange multipliers $\mathcal{A}, \mathcal{B}$ disappear, since our definitions for $\mathbf{R}$ and π are already normalised. The maximum of $\mathcal{L}$ is $g = \hat{S}/(\hat{S}+1)$, $\lambda = \hat{I}/(\hat{Z}+T)$, $\mu = \hat{D}/\hat{Z}$.

The supplementary information describes a procedure for calculating $\hat{S}$, $\hat{Z}$, $\hat{D}$ and $\hat{I}$ from sequence data.

4 RESULTS

4.1 Simulated data

We implemented the above-described EM algorithm for estimating indel rates in under 300 lines of C. The implemented code is available free of charge under the GNU Public License from the website `www.biowiki.org`. (The related EM algorithm for estimating substitution rates was previously implemented in the `xrate` program, available from the same website [7].)

To test the EM algorithm, we ran extensive numerical simulations, generating random pairwise alignments under the TKF model and comparing the expected insertion counts ($\hat{I}$) and site-times ($\hat{Z}$) with the actual values of these counts, known from the simulation. We also looked at the following "naive" estimates of these quantities: for I, we simply counted the number of insertions in the simulated alignments, while for Z, we multiplied the lapsed time T by the average of the ancestor and descendant sequence lengths.

Figure 2 shows the results of these simulations. There is extremely close agreement between the computed expectations ($\hat{I}$, $\hat{Z}$) and the actual values of these statistics. As for the "naive" estimates, we observe that simply counting the number of insertions in the pairwise alignment is a systematically biased under-estimate of I, as expected (because some insertions were deleted before being observed). However, perhaps surprisingly, the naive estimate of Z (based on the average of the ancestor and descendant sequence lengths) turned out to be rather good. On closer investigation, we found that the naive Z estimate deteriorates significantly when λ is much less than μ, and also appears to perform worse at larger timescales. In practise, for long sequences (so $\lambda \simeq \mu$) that are closely related (so T is small), the naive estimator for Z may be appropriate, although it seems likely that I and D will still be underestimates.

4.2 Biological data

To test the application of the indel EM algorithm to experimental data, we estimated the indel rates for multiple alignments of Coronavirus protein domains, including domains from the SARS coronavirus [20]. Lengths of the SARS domains in amino acids (aa) are reported below. The first domain is GS1 (629aa) and the second is GS2 (626aa); both

are derived from peptide cleavage of the "Spike" surface glycoprotein. The third domain is C16 (277aa), the Papain-like peptidase domain from the long RNA polymerase gene product. The analysed proteins were sequenced from viruses responsible for SARS, murine hepatitis, porcine transmissible gastroenteritis, porcine epidemic diarrhea, equine arteritis and avian infectious bronchitis.

Indel rates were estimated using three different methods: the EM method presented here, a Markov Chain Monte Carlo (MCMC) method and a naive estimate. The MCMC runs included 10^5 datapoints constituting $\sim 10^3$ effective independent samples. Confidence intervals for the MCMC-estimated rate parameters are reported as $m \pm 2s$ where m is the mean and s the standard deviation of the marginal posterior distribution over the relevant parameter. The EM and naive methods do not return error estimates.

The domain boundaries were first identified by reference to the Pfam database [21]. We then estimated a separate phylogenetic tree for each domain, first estimating a pairwise distance matrix using a 20×20 amino acid substitution rate matrix previously estimated from PFAM alignments by the EM algorithm [7], then finding an approximate phylogenetic tree using weighted neighbor-joining [22]. The three trees thus estimated are shown in Figures 3-5.

Next, we used the `tkfalign` program [6] to infer maximum likelihood alignment paths for the missing ancestral sequences, using that program's default values for the indel rates ($\lambda = 0.049505$, $\mu = 0.050495$). We used the alignments containing these inferred ancestral sequences in all subsequent analyses. This strategy is likely to systematically under-estimate the true indel rates, since the ML alignment paths will tend to minimize the number of inferred events. Ideally, we should integrate out the ancestral sequences, e.g. by Monte Carlo alignment sampling [6] or multidimensional dynamic programming [5]. However, fixing the alignments simplifies the comparison of the parameter estimation methods.

We proceeded to estimate the indel rates, using the three methods described (naive, EM and MCMC). The results of this analysis are shown in Table 1. We note that the Spike protein has elevated indel rates compared to the peptidase protein, and that the first (5') domain has higher rates than the second (3') domain. Some caution is, however, needed in interpreting these results, because the indel rate estimates depend strongly on the estimated tree. Since we have estimated independent phylogenies (including branch lengths) for the GS1 and GS2 domains, and since the branch length estimates are dominated by the substitutions in the alignment, what we are actually measuring here is not the absolute indel rate but the ratio of the indel rate to the substitution rate. If substitutions are generally occurring faster in one of the domains (as indeed they are, in the GS1 domain) this may significantly skew the estimate.

An absolute comparison of the rates of evolution of the Spike domains can be made by the following procedure. First, estimate a phylogenetic tree for the full alignment of the entire Spike protein (Figure 6); next, estimate λ and μ independently for the two domains, using the same tree (and branch lengths) for each domain. The results of this test are given in Table 2. Here, it can be seen that the "absolute" rate of indel evolution is, in fact, greater in GS1. Thus, the overall picture is that GS1 is mutating faster than GS2, but substitutions are elevated relative to indels. Given the relative mutation rates, we predict that the GS1 domain has more contact interactions with the host immune system than GS2.

We note that although there is a clear systematic bias to the naive rate estimates, this bias ranges from $2\% \sim 9\%$ and, as such, is usually slightly smaller than the sampling error (as revealed by MCMC). The exception to this is the GS1 domain, for which the bias is slightly larger (but not by much). These results seem to indicate that the naive estimates would probably be acceptable in most practical situations, although in order to avoid the possibility of downstream or compounded errors (e.g. in more elaborate TKF91-derived models containing more parameters [23]) it would be better to use the unbiased estimators derived here.

5 DISCUSSION

We have shown that the expected initial sequence length ($\hat{S}$), numbers of mortal insertions ($\hat{I}$) and deletions ($\hat{D}$), and accrued time for mortal ($\hat{Z}$) and immortal (T) links, along with the expected initial composition ($\hat{s}_i$), substitution counts ($\hat{u}_{ij}$) and accrued times ($\hat{w}_i$) for each residue, constitute "sufficient statistics" for maximising the expected log-likelihood of a given set of alignments under the TKF91 evolutionary model. This EM algorithm is orders of magnitude faster than brute-force methods of searching the indel/substitution rate space, requiring only an $O((T/\Delta t)^2)$ overhead for numerical integration, where Δt is the timestep for the discretized integral. We have implemented such an EM algorithm and tested it by simulation and application to biological data.

In the prevalent situation where a multiple alignment is unknown, the calculation of the sufficient statistics by summing over all alignments takes time $O(L^N)$, where L is the (geometric mean) sequence length and N is the number of sequences. However, a stochastic version of the EM algorithm, which uses MCMC alignment sampling on local neighbourhoods of the tree, can approximate these statistics efficiently (and improve alignment accuracy) in $O(SL^K)$ time, where K is the neighbourhood size (i.e. the number of nodes whose mutual alignment is simultaneously resampled at any one step) and S is the number of sampling steps between stochastic EM updates. Naturally, one can also proceed on the assumption that the best alignment found by an alignment algorithm is the "true" multiple alignment, though

such an assumption may lead to systematic biases such as undercounting indels.

Recent years have seen considerable interest in the derivation of stochastic evolutionary processes for biological sequences and/or structures [18, 19, 9, 24, 25, 23]. Several of these are so closely related to TKF91 that the theory derived here is directly applicable. For example, the TKF92 model [18] (see also Section 2.3) is a birth-death process on sequence fragments, rather than residues, so that the number of matches (a quantity appearing in the expressions for the EM-sufficient statistics) is found by counting the number of aligned fragments rather than the number of aligned residues. Similarly, an evolutionary model for RNA secondary structure has recently been described, together with a recipe for computing the sufficient statistics using the Inside-Outside algorithm [23]. Models that depart more substantially from TKF91, such as "long indel" models [19, 9] or context-sensitive substitution models [24, 25], have generally not been solved as completely (i.e. we still lack exact algebraic expressions for the sequence likelihood) and so the TKF91 theory developed here will be less directly applicable. These models may mandate different approaches to rate estimation, such as enumeration of mutation trajectories [9], truncated Taylor expansions [24] or variational approximations [25].

The EM algorithm—in the form of the Baum-Welch algorithm—is one cornerstone of the application of profile HMMs in bioinformatics. We hope that the theoretical developments described here may contribute to the efforts to make evolutionary models similarly useful for probabilistic sequence analysis and phylogenomics.

λ (insertion rate)

Domain	Naive	**EM**	MCMC
GS1	0.135352	**0.145626**	0.145 ± 0.008
GS2	0.0928362	**0.0971039**	0.098 ± 0.008
C16	0.0341402	**0.0346793**	0.035 ± 0.009

μ (deletion rate)

Domain	Naive	**EM**	MCMC
GS1	0.135459	**0.145741**	0.146 ± 0.008
GS2	0.0929047	**0.0971758**	0.098 ± 0.008
C16	0.0342055	**0.0347457**	0.035 ± 0.009

Table 1. Indel rates for coronavirus protein domains. A separate phylogenetic tree was estimated for each domain (Figures 3-5). Key to domain names: GS1 is the 5' domain of the SARS coronavirus Spike surface glycoprotein; GS2 is the 3' domain of the same protein; and C16 is the peptidase domain of the SARS virus replicase protein.

λ (insertion rate)

Domain	Naive	**EM**	MCMC
GS1	0.105146	**0.110403**	0.110 ± 0.006
GS2	0.0514705	**0.0527830**	0.053 ± 0.006

μ (deletion rate)

Domain	Naive	**EM**	MCMC
GS1	0.105219	**0.110480**	0.111 ± 0.006
GS2	0.0515080	**0.0528216**	0.053 ± 0.006

Table 2. Indel rates for coronavirus Spike protein domains. The same phylogenetic tree was used for both domains (Figure 6). Key to domain names: GS1 is the 5' domain of the SARS coronavirus Spike surface glycoprotein, while GS2 is the 3' domain of the same protein.

ACKNOWLEDGEMENTS

This work was partially supported by grants from EPSRC (code HAMJW) and MRC (code HAMKA). The author would like to thank Bob Griffiths, Von Bing Yap and Terry Speed for helpful discussions, and two anonymous reviewers for their suggestions.

REFERENCES

[1]R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK, 1998.

[2]A. P Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38, 1977.

[3]G. J. Mitchison and R. Durbin. Tree-based maximal likelihood substitution matrices and hidden Markov models. *Journal of Molecular Evolution*, 41:1139–1151, 1995.

[4]J. Hein, C. Wiuf, B. Knudsen, M. B. Moller, and G. Wibling. Statistical alignment: computational properties, homology testing and goodness-of-fit. *Journal of Molecular Biology*, 302:265–279, 2000.

[5]J. Hein. An algorithm for statistical alignment of sequences related by a binary tree. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, pages 179–190, Singapore, 2001. World Scientific.

[6]I. Holmes and W. J. Bruno. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, 17(9):803–820, 2001.

[7]I. Holmes and G. M. Rubin. An Expectation Maximization algorithm for training hidden substitution models. *Journal of Molecular Biology*, 317(5):757–768, 2002.

[8]I. Holmes. Using guide trees to construct multiple-sequence evolutionary HMMs. In *Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology*, pages 147–157, Menlo Park, CA, 2003. AAAI Press.

[9]I. Miklós, G. Lunter, and I. Holmes. A long indel model for evolutionary sequence alignment. *Molecular Biology and Evolution*, 21(3):529–540, 2004.

[10]J. L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, 33:114–124, 1991.

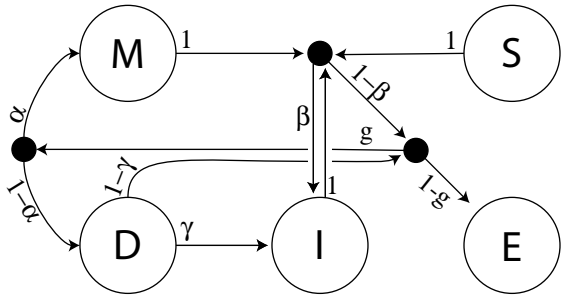

Fig. 1. A Pairwise Hidden Markov Model (Pair HMM) representation of the TKF91 evolutionary model. Transition probabilities are shown beside each transition arrow. The model contains **S**tart, **E**nd, **M**atch, **D**elete and **I**nsert states, plus three null states (filled black circles) introduced to simplify the transition probability expressions. (Note that Start and End may also be considered null states.)

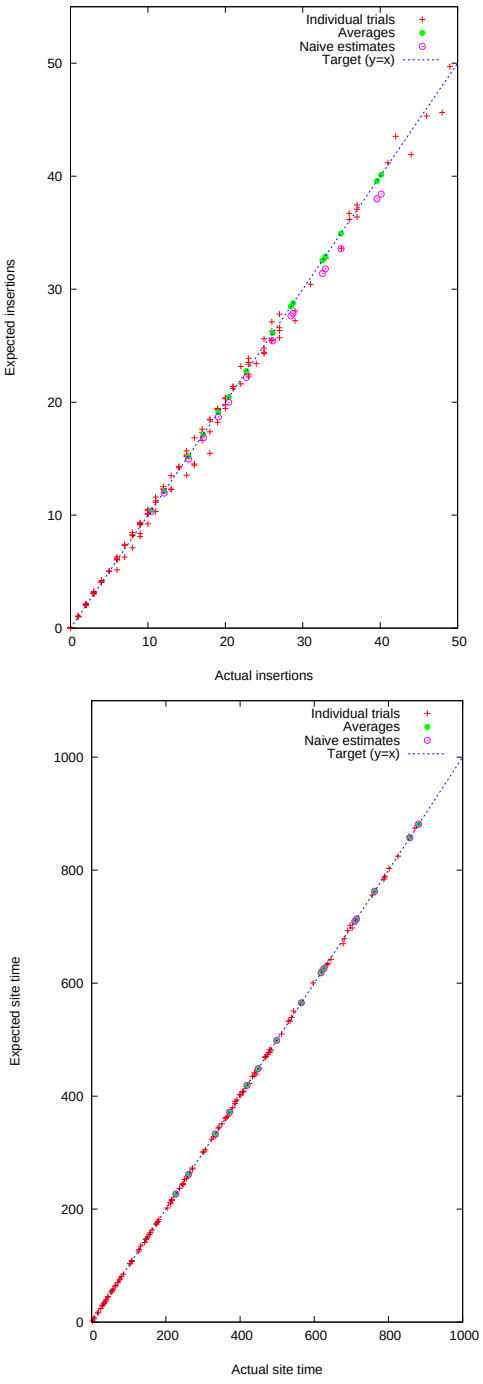

Fig. 2. Estimated vs actual results for number of insertions (left) and accrued site-time (right) obtained by simulation with $\lambda = .0457$, $\mu = .0458$ (mean sequence length $\simeq 50$ residues), T ranging from 0.5 to 2, 10^5 trials per timepoint. The average over each 10^5 trials is plotted; to give an impression of error on the individual estimates, one in every 5×10^4 individual trials is also plotted. The "naive" estimates, taken without correcting for statistical bias, are also shown. The target values are shown as a dashed straight line.

[11]D. Metzler, R. Fleissner, A. von Haeseler, and A. Wakolbinger. Assessing variability by joint sampling of alignments and mutation rates. *Journal of Molecular Evolution*, 53(6):660–669, 2001.

[12]N. Friedman, M. Ninio, I. Pe'er, and T. Pupko. A structural EM algorithm for phylogenetic inference. In T. Lengauer, D. Sankoff, S. Istrail, P. Pevzner, and M. Waterman, editors, *Proceedings of the Fifth Annual International Conference on Computational Biology*, New York, 2001. Association for Computing Machinery.

[13]J. L. Thorne, N. Goldman, and D. T. Jones. Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, 13:666–673, 1996.

[14]M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, 2003.

[15]J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc., 2003. ISBN 0878931775.

[16]W. H. Press, S. A. Teukolsky, W. Vetterling T., and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK, 1992.

[17]S. Karlin and H. Taylor. *A First Course in Stochastic Processes*. Academic Press, San Diego, CA, 1975.

[18]J. L. Thorne, H. Kishino, and J. Felsenstein. Inching toward reality: an improved likelihood model of sequence evolution. *Journal of Molecular Evolution*, 34:3–16, 1992.

[19]B. Knudsen and M. Miyamoto. Sequence alignments and pair hidden Markov models using evolutionary history. *Journal of Molecular Biology*, 333(2):453–460, 2003.

[20]M. A. Marra, S. J. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. Butterfield, J. Khattra, J. K. Asano, S. A. Barber, S. Y. Chan, A. Cloutier, S. M. Coughlin, D. Freeman, N. Girn, O. L. Griffith, S. R. Leach, M. Mayo, H. McDonald, S. B. Montgomery, P. K. Pandoh, A. S. Petrescu, A. G. Robertson, J. E. Schein, A. Siddiqui, D. E. Smailus, J. M. Stott, G. S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T. F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando,

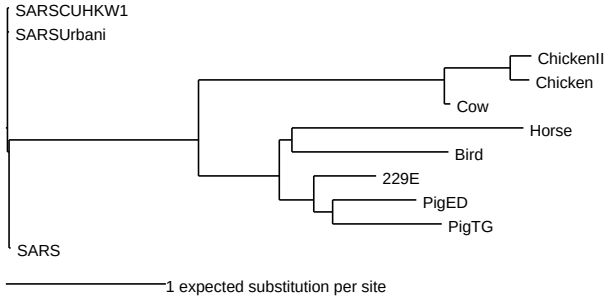

Fig. 3. Phylogenetic trees estimated by the neighbor-joining method applied to pairwise amino acid-level distance matrices for the 5' GS1 domain from the coronavirus glycosurfaceprotein "Spike". Key: Chicken=AAF69334.1 (Murine hepatitis virus), ChickenII=AAF19386.1 (Murine hepatitis virus strain II), Cow=NP_150077.1 (Bovine coronavirus), SARS=NP_828851.1 (SARS coronavirus), SARSCUHKW1=AAP13567.1 (SARS strain CUHK-W1), SARSUrbani=AAP13441.1 (SARS strain Urbani), Bird=NP_740621.1 (Avian infectious bronchitis virus), 229E=NP_073551.1 (Human coronavirus strain 229E), PigED=NP_598310.1 (Porcine epidemic diarrhea virus), PigTG=NP_058424.1 (Porcine transmissible gastroenteritis virus), Horse=NP_705583.1 (Equine arteritis virus).

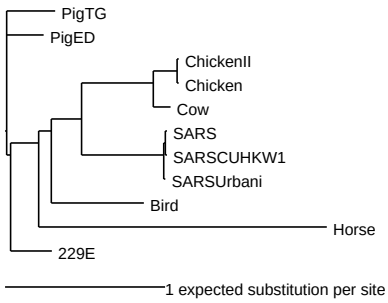

Fig. 4. Phylogenetic trees estimated by the neighbor-joining method applied to pairwise amino acid-level distance matrices for the 3' GS2 domain from the coronavirus glycosurfaceprotein "Spike". For sequence accession numbers, see legend to Figure 3.

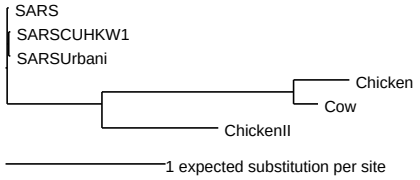

Fig. 5. Phylogenetic trees estimated by the neighbor-joining method applied to pairwise amino acid-level distance matrices for the Papain-like peptidase domain from coronavirus RNA polymerase. Key: Chicken=AAF69341.1, ChickenII=AAF19383.1, Cow=NP_742129.1, SARS=NP_828860.1, SARSUrbani=AAP13439.1, SARSCUHKW1=AAP13575.1.

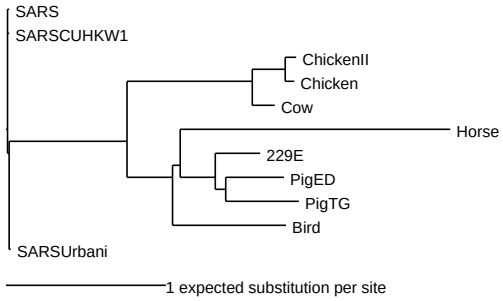

Fig. 6. Phylogenetic trees estimated by the neighbor-joining method applied to pairwise amino acid-level distance matrices for the entire "Spike" coronavirus glycosurfaceprotein (including both domains GS1 and GS2). For sequence accession numbers, see legend to Figure 3.

R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G. A. Tipples, S. Tyler, R. Vogrig, D. Ward, B. Watson, R. C. Brunham, M. Krajden, M. Petric, D. M. Skowronski, C. Upton, and R. L. Roper. The genome sequence of the SARS-associated coronavirus. *Science*, 300(5624):1399–1404, 2003.

[21]A. Bateman, E. Birney, R. Durbin, S. R. Eddy, R. D. Finn, and E. L. Sonnhammer. Pfam 3.1: 1313 multiple alignments match the majority of proteins. *Nucleic Acids Research*, 27(1):260–262, 1999.

[22]W. J. Bruno, N. D. Socci, and A. L. Halpern. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution*, 17(1):189–197, 2000.

[23]I. Holmes. A probabilistic model for the evolution of RNA structure. *BMC Bioinformatics*, 5(166), 2004.

[24]G.A. Lunter and J. Hein. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, 2004. To appear.

[25]A. Siepel and D. Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution*, 21(3):468–488, 2004.

[26]J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[27]S. Karlin and J. McGregor. The classification of birth and death processes. *Transactions of the American Mathematical Society*, 86:366–400, 1957.

[28]S. Karlin and J. McGregor. Linear growth, birth and death processes. *Journal of Mathematics and Mechanics*, 7:643–662, 1958.