Phylogenetics

Using evolutionary Expectation Maximization to estimate indel rates

Ian Holmes*

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

Received on July 20, 2004; revised on November 19, 2004; accepted on November 22, 2004
Advance Access publication February 24, 2005

ABSTRACT
Motivation: The Expectation Maximization (EM) algorithm, in the form of the Baum–Welch algorithm (for hidden Markov models) or the Inside-Outside algorithm (for stochastic context-free grammars), is a powerful way to estimate the parameters of stochastic grammars for biological sequence analysis. To use this algorithm for multiple-sequence evolutionary modelling, it would be useful to apply the EM algorithm to estimate not only 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 EM, under the single-residue indel model owing 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). Owing 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: ihh@berkeley.edu

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 (Durbin et al., 1998); both are examples of the Expectation Maximization (EM) algorithm (Dempster et al., 1977).

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 (Durbin et al., 1998)]. 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 arbitrary 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 (Mitchison and Durbin, 1995; Hein et al., 2000; Hein, 2001; Holmes and Bruno, 2001; Holmes and Rubin, 2002; Holmes, 2003; Miklós et al., 2004). 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 (ML) rate matrices for point substitution processes (Holmes and Rubin, 2002). 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 (1) the expected composition of the root sequence \( \hat{s} \), (2) the expected number of times, \( \hat{u}_{ij} \), that each mutation of type \( i \rightarrow j \) occurred and (3) the expected length of time, \( \hat{v}_i \), 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 \( \hat{s} \) and substitution rates equal to \( \hat{u}_{ij}/\hat{v}_i \). This process converges extremely rapidly and, assuming the absence of significantly suboptimal local minima in the likelihood function (which in practice is often the case), results in ML substitution rate matrices.

This paper presents an EM algorithm for the estimation of indel rates in the TKF91 model of sequence evolution (Thorne et al., 1991). The main insight is that the analogues of the sufficient statistics (1) \( \hat{s} \), (2) \( \hat{u}_{ij} \) and (3) \( \hat{v}_i \) for the TKF91 indel process are (1) the expected initial sequence length, (2) the expected number of insertions and deletions and (3) 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
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: (1) point substitutions, which act on a single residue only (this process uses the finite-state Markov chain framework described in the previous section); and (2) single-residue indels, which insert or delete a single residue (Thorne et al., 1991). Insertions occur at rate \( \lambda \) per available site; deletions at rate \( \mu \).

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 (Karlin and Taylor, 1975). The rate matrix is sparse: \( R_j = (i + 1)\lambda \) if \( j = i + 1 \), \( R_{ij} = i\mu \) if \( j = i - 1 \) and \( R_{ij} = 0 \) if \(|i - j| > 1\). Thus \( R_j = -(i+1)\lambda - i\mu \). 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 \( n \); and second (for mortal links) by specifying whether the original ancestral residue survived or was deleted. Let the probability distribution of \( n \) be \( p_n(t) \) for immortal links and \( q_n(t) \) 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 (Thorne et al., 1991) that

\[
\begin{align*}
p_n(t) &= \alpha(t)\beta(t)^{n-1}(1 - \beta(t)) \\
q_n(t) &= (1 - \alpha(t))(1 - \gamma(t)) \quad \text{for } n = 0 \\
r_n(t) &= \beta(t)^n(1 - \beta(t)) \quad \text{for } n > 0
\end{align*}
\]

where

\[
\begin{align*}
\alpha(t) &= e^{-\mu t} \\
\beta(t) &= \frac{\lambda(1 - e^{(\mu-\lambda)t})}{\mu - \lambda e^{(\mu-\lambda)t}} \\
\gamma(t) &= 1 - \frac{\mu(1 - e^{(\mu-\lambda)t})}{(1 - e^{-\mu t})(\mu - \lambda e^{(\mu-\lambda)t})}
\end{align*}
\]

This probabilistic model for pairwise alignments can be expressed as a pairwise HMM using, e.g. the notation of Durbin et al. (1998), as shown in Figure 1.

The restriction to single-residue indel events leads to the geometic term \( \beta^n \) 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 (Thorne et al., 1992; Hein et al., 2000; Knudsen and Miyamoto, 2003; Miklós et al., 2004). 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 (Knudsen and Miyamoto, 2003; Miklós et al., 2004). 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 a general theory for continuous-time Markov models, and then 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 $h$ 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 $O$ which partially constrains the allowable histories $h$. 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 (Dempster et al., 1977) consists of maximizing the following sum over possible histories with respect to $\theta'$:

$$Q(\theta, \theta') = \sum_h P(h|O, T, \theta) \log P(h|O|T, \theta').$$

Previous work (Holmes and Rubin, 2002) showed that, if we break the time interval $[0, T]$ into small intervals $\Delta t$, we obtain

$$Q(\theta, \theta') = \sum_i \hat{s}_i \log \pi_i + \sum_i \hat{w}_i R_{ii}^T + \sum_i \sum_{j \neq i} \hat{u}_{ij} \log R_{ij}^T + f(\theta, \Delta t),$$

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, $f(\theta, \Delta t)$ is a function that does not depend on the ‘new’ parameters $\theta'$. In fact $f(\theta, \Delta t) \sim -\log(\Delta t)$.

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

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

Introducing these constraints via Lagrange multipliers $\{A_1, \ldots, A_m, B\}$ and dropping $f$, the function to be maximized is

$$L(\theta, \theta', \alpha, \beta) = \sum_i \hat{s}_i \log \pi_i + \sum_i \hat{w}_i R_{ii}^T + \sum_i \sum_{j \neq i} \hat{u}_{ij} \log R_{ij}^T + B \left( \sum_i \pi_i^T - 1 \right).$$

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

$$\hat{s}_i(a, b, T) = \delta_{ia},$$

$$\hat{w}_i(a, b, T) = \frac{T_{ji}^{ab}(T)}{M_{ji}(T)},$$

$$\hat{u}_{ij}(a, b, T) = \frac{T_{ij}^{ab}(T)}{M_{ij}(T)} R_{ij},$$

where

$$T_{ij}^{ab}(T) = \int_0^T M_{ij}(t) M_{ji}(T - t) \, dt = \left( \frac{\pi_i \pi_j}{\pi_a \pi_j} \right)^{1/2} \sum_k \sum_l \frac{v_i^{(k)}}{v_j^{(l)}} \frac{v_j^{(l)}}{v_i^{(k)}} J_{kl}(T)$$

with

$$J_{kl}(T) = \begin{cases} T(\chi^{(k)} - \chi^{(l)}) + \frac{\chi^{(k)}}{\mu^{(k)} - \mu^{(l)}} (\mu^{(k)} - \mu^{(l)}) & \text{if } \mu^{(k)} = \mu^{(l)} \\
\chi^{(k)} - \chi^{(l)} & \text{if } \mu^{(k)} \neq \mu^{(l)}.\end{cases}$$

The maximum of $L$ 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 $R_{ij}'' = -\sum_j R_{ij}'$.
For the equilibrium distribution $\pi$, use independent geometric length parameter $g$, so that $\pi_i = g^{(i-1)}(1-g)$ (as long as the process stays reversible, we will end up with $g = \lambda/\mu$).

$$\mathcal{L}(\theta, \theta') = \sum_{i=0}^{\infty} \hat{S}_i \left( i \log g + \log(1-g) \right)$$

$$- \sum_{i=0}^{\infty} \hat{w}_i \left( (i+1) \lambda + i \mu \right)$$

$$+ \sum_{i=0}^{\infty} \hat{u}_{i,i-1} \log(i \mu')$$

$$+ \sum_{i=0}^{\infty} \hat{u}_{i,i+1} \log((i+1) \lambda')$$

$$= \hat{S} \log g + \log(1-g) - (\hat{Z} + T) \lambda' - \hat{Z} \mu' + \hat{D} \log \mu' + \hat{I} \log \lambda' + f'(\theta),$$

where $\hat{S} = \sum_i \hat{S}_i$ 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 $f'(\theta)$ does not depend on $\theta'$.

Note the Lagrange multipliers $A, B$ disappear, since our definitions for $R$ and $\pi$ are already normalized. 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 (Holmes and Rubin, 2002).]

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 underestimate 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 $\lambda$ is much less than $\mu$, and also appears to perform worse at larger timescales. In practice, for
etic tree using weighted neighbour-joining (Bruno et al., 2002), then finding an approximate phylogenetic tree for each domain, first estimating a pairwise distance matrix using a 20 amino acid substitution rate matrix previously estimated from PFAM alignments by the EM algorithm (Marra et al., 2003). Lengths of the SARS domains in amino acids are reported below. The first domain is GS1 (629 amino acids) and the second is GS2 (626 amino acids); both are derived from peptide cleavage of the ‘Spike’ surface glycoprotein. The third domain is C16 (277 amino acids), 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 diarrhoea, 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 $10^5$ 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 (Bateman et al., 1999). We then estimated a separate phylogenetic tree for each domain, first estimating a pairwise distance matrix using a $20 \times 20$ amino acid substitution rate matrix previously estimated from Pfam alignments by the EM algorithm (Holmes and Rubin, 2002), then finding an approximate phylogenetic tree using weighted neighbour-joining (Bruno et al., 2000). The three trees thus estimated are shown in Figures 3–5.

Next, we used the tklalign program (Holmes and Bruno, 2001) to infer ML 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 underestimate the true indel rates, since the ML alignment paths will tend to minimize the number of inferred events.

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 (Marra et al., 2003). Lengths of the SARS domains in amino acids are reported below. The first domain is GS1 (629 amino acids) and the second is GS2 (626 amino acids); both are derived from peptide cleavage of the ‘Spike’ surface glycoprotein. The third domain is C16 (277 amino acids), 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 diarrhoea, 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 $10^5$ 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 (Bateman et al., 1999). We then estimated a separate phylogenetic tree for each domain, first estimating a pairwise distance matrix using a $20 \times 20$ amino acid substitution rate matrix previously estimated from Pfam alignments by the EM algorithm (Holmes and Rubin, 2002), then finding an approximate phylogenetic tree using weighted neighbour-joining (Bruno et al., 2000). The three trees thus estimated are shown in Figures 3–5.

Next, we used the tklalign program (Holmes and Bruno, 2001) to infer ML 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 underestimate 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 (Holmes and Bruno, 2001) or multidimensional dynamic programming (Hein, 2001). 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 with 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 (Fig. 6); next, estimate \( \lambda \) and \( \mu \) 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 to \( \sim9\% \) 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 (Holmes, 2004)] it would be better to use the unbiased estimators derived here.

### Table 1. Indel rates for coronavirus protein domains

| Domain | Naive         | EM           | MCMC        |
|--------|---------------|--------------|-------------|
| GS1    | 0.135352      | 0.145626     | 0.145 ± 0.008 |
| GS2    | 0.0929362     | 0.0971039    | 0.098 ± 0.008 |
| C16    | 0.0341402     | 0.0346793    | 0.035 ± 0.009 |

### Table 2. Indel rates for coronavirus Spike protein domains

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

#### 5 DISCUSSION

We have shown that the expected initial sequence length (\( \hat{S} \)), numbers of mortal insertions (\( I \)) and deletions (\( D \)), and accrued time for mortal (\( \hat{Z} \)) and immortal (\( \hat{T} \)) links, along with the expected initial composition (\( \hat{c} \)), substitution counts (\( u_{ij} \)) and accrued times (\( \hat{w}_i \)) for each residue, constitute ‘sufficient statistics’ for maximizing 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 \( \Delta t \) 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 \( O(SL^2) \) 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, although 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 (Thorne et al., 1992; Knudsen and Miyamoto, 2003; Miklós et al., 2004; Lunter and Hein, 2004; Siepel and Haussler, 2004; Holmes, 2004). Several of these are so closely related to TKF91 that the theory derived here is directly applicable. For example, the TKF92 model (Thorne et al., 1992) (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 (Holmes, 2004). Models that depart more substantially from TKF91, such as ‘long indel’ models (Knudsen and Miyamoto, 2003; Miklós et al., 2004) and...
context-sensitive substitution models (Lunter and Hein, 2004; Siepel and Haussler, 2004), have generally not been solved 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 (Miklós et al., 2004), truncated Taylor expansions (Lunter and Hein, 2004) and variational approximations (Siepel and Haussler, 2004).

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.

ACKNOWLEDGEMENTS

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

SUPPLEMENTARY DATA

Supplementary data for this paper are available on Bioinformatics online.

REFERENCES

Bateman,A. et al. (1999) Pfam 3.1: 1313 multiple alignments match the majority of proteins. Nucleic Acids Res., 27, 260–262.

Bruno,W.J. et al. (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol. Biol. Evol., 17, 189–197.

Dempster,A.P. et al. (1977) Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., B, 39, 1–38.

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

Felsenstein,J. (2003)Inferring Phylogenies. Sinauer Associates, Inc., ISBN 0878931775.

Friedman,N., Ninio,M., Pcer,I. and Popko,T. (2001) A structural EM algorithm for phylogenetic inference. In Lengauer,T., Sankoff,D., Istrail,S., Pevzner,P. and Waterman,M. (eds), Proceedings of the Fifth Annual International Conference on Computational Biology, New York. Association for Computing Machinery. Hein,J. (2001) An algorithm for statistical alignment of sequences related by a binary tree. In Altman,R.B., Dunker,A.K., Hunter,L., Lauderdale,K. and Klein,T.E. (eds), Pacific Symposium on Biocomputing. World Scientific, Singapore, 179–190.

Hein,J. et al. (2000) Statistical alignment: computational properties, homology testing and goodness-of-fit. J. Mol. Biol., 302, 265–279.

Holmes,J. (2003) Using guide trees to construct multiple-sequence evolutionary HMMs. In Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology, Menlo Park, CA. AAAI Press, pp. 147–157.

Holmes,J. (2004) A probabilistic model for the evolution of RNA structure. BMC Bioinformatics, 5.

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

Holmes,J. and Rubin,G.M. (2002) An Expectation Maximization algorithm for training hidden substitution models. J. Mol. Biol., 317, 757–768.

Karlin,S. and Taylor,H. (1975) A First Course in Stochastic Processes. Academic Press, San Diego, CA.

Kelley,M. et al. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature, 423, 241–254.

Knudsen,B. and Miyamoto,M. (2003) Sequence alignments and pair hidden Markov models using evolutionary history. J. Mol. Biol., 333, 453–460.

Lunter,G.A. and Hein,J. (2004) A nucleotide substitution model with nearest-neighbour interactions. Bioinformatics, 20(Suppl. 1), i216–i223.

Marz,M.A. et al. (2003) The genome sequence of the SARS-associated coronavirus. Science, 300, 1399–1404.

Metzler,D. et al. (2001) Assessing variability by joint sampling of alignments and mutation rates. J. Mol. Evol., 53, 660–669.

Miklós,I. et al. (2004) A long indel model for evolutionary sequence alignment. Mol. Biol. Evol., 21, 529–540.

Mitchison,G.J. and Durbin,R. (1995) Tree-based maximal likelihood substitution matrices and hidden Markov models. J. Mol. Biol., 241, 1139–1151.

Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (1992) Numerical Recipes in C. Cambridge University Press, Cambridge, UK.

Siepel,A. and Haussler,D. (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol., 21, 468–488.

Thorne,J.L. et al. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol., 33, 114–124.

Thorne,J.L. et al. (1992) Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol., 34, 3–16.

Thorne,J.L. et al. (1996) Combining protein evolution and secondary structure. Mol. Biol. Evol., 13, 666–673.