Applications of Hidden Markov Models for Characterization of Homologous DNA Sequences with a Common Gene

ASGER HOBOLTH\textsuperscript{1} and JENS LEDET JENSEN\textsuperscript{2}

ABSTRACT

Identifying and characterizing the structure in genome sequences is one of the principal challenges in modern molecular biology, and comparative genomics offers a powerful tool. In this paper, we introduce a hidden Markov model that allows a comparative analysis of multiple sequences related by a phylogenetic tree, and we present an efficient method for estimating the parameters of the model. The model integrates structure prediction methods for one sequence, statistical multiple alignment methods, and phylogenetic information. This unified model is particularly useful for a detailed characterization of DNA sequences with a common gene. We illustrate the model on a variety of homologous sequences.

Key words: alignment, comparative genomics, EM-algorithm, gene finding, hidden Markov model, phylogeny, structure prediction.

1. INTRODUCTION

Structure identification of genome sequences is a central challenge in molecular biology. Comparative genomics provide a powerful and general approach for identifying functional elements such as genes. Natural selection implies that functional elements should have a larger degree of conservation across related species than elements with no function. The power of comparative genomics increases with the number of species, and therefore the approach is likely to become increasingly important as more genomes are being sequenced. The main purpose of this paper is to develop and apply statistical approaches for systematic analysis of several related genomic sequences.

Hidden Markov models (HMMs) along the sequence have been successfully applied to gene structure prediction in one sequence (cf., e.g., Burge and Karlin [1997] and Krogh [1997]). The one-sequence HMMs partition a sequence into (at least) five parts: one part representing the sequence before the gene, one representing the start of the gene, one representing the inside of the gene, one representing the stop of the gene, and one representing the sequence after the gene. If the sequence is from an eukaryotic organism, the part of the sequence inside the gene is further divided into alternating coding and noncoding parts (exons and introns).

Recently, Korf et al. (2001), Pachter et al. (2002), and Meyer and Durbin (2002) have extended the gene-structure-prediction HMMs for one sequence to two sequences. Their HMMs simultaneously predict
the gene structure and align two homologous sequences, and the transition and substitution probabilities of the models are determined from training data. In order to extend the pair HMMs for simultaneous gene structure prediction and alignment to multiple sequences, the transition and substitution probabilities should be derived from the evolutionary relationship between the sequences. We propose a model that integrates gene structure prediction, alignment methods, and phylogenetic information. The model is fully parametric and can in principle be extended to any number of homologous sequences. Further, we develop a novel method for parameter estimation based on the expectation maximization (EM) algorithm and moment equations. The proposed HMM provides a detailed characterization of homologous DNA sequences with one common gene. For example, the model includes key parameters such as branch lengths, transition–transversion ratios, and synonymous–nonsynonymous ratios. The model also provides insight into the relative likelihood of, e.g., starting or ending a gene at certain places in a sequence. Thus, the model developed in this paper is useful for close investigation of multiple small homologous sequences.

In the before-gene, after-gene, and intronic parts, we assume that the evolution from one sequence to the other follows the Thorne, Kishino, and Felsenstein (1991) model. If an ancestral sequence $S_1$ has evolved to a sequence $S_2$, the evolution can be summarized in terms of an alignment of some of the letters in $S_1$ with some of the letters in $S_2$, in terms of deletions of some of the letters in $S_1$, in terms of insertions of some of the letters in $S_2$, and in terms of substitutions of the aligned letters. The TKF-model can be formulated as a hidden Markov model (HMM) along the sequence with three hidden states corresponding to match (a pair of aligned letters), deletion, or insertion of single nucleotides. The substitution probabilities are determined by the Hasegawa, Kishino and Yano (1985) model, which involves parameters describing nucleotide frequencies and the transition to transversion ratio.

In the coding part of the gene, the sequences have also evolved according to the TKF-model, but formulated on the codon level. Thus, the hidden states correspond to match, deletion, and insertion of nucleotide triplets. The substitution probabilities are determined by the codon model of Goldman and Yang (1994). Besides parameters of codon frequencies and the transition to transversion ratio, the codon model also distinguishes between nonsynonymous and synonymous codon substitutions. The start, stop, donor, and acceptor site positions are modeled in terms of simple functional signals.

Pedersen and Hein (2003) also predict gene structure in multiple related sequences, but their HMM assumes that an alignment of the sequences has already been established. In this paper, we extend Pedersen and Hein (2003) to perform gene finding and alignment simultaneously and Pachter et al. (2002) and Meyer and Durbin (2002) to treat more than two sequences.

In the three following sections of this paper, we consider pairwise prokaryotic, pairwise eukaryotic, and triplewise prokaryotic gene structure prediction. The parametric hidden Markov models are described in detail, and an EM-algorithm for parameter estimation is developed. We also apply the suggested models to DNA sequence data. The paper concludes with a discussion of extensions of the models to more than three species.

## 2. PAIRWISE PROKARYOTIC GENE STRUCTURE PREDICTION

Let $S_1$ and $S_2$ denote two observed homologous DNA sequences of lengths $L_1$ and $L_2$ from prokaryotic organisms. The $i$th nucleotide in sequence $j$ is $S_j[i]$. We use a hidden Markov model (HMM) along the sequences to describe the evolutionary relationship of the two sequences. A HMM consists of a set of hidden states that determine the underlying (hidden) structure of the sequences. If the sequences contain one common gene, the hidden state sequence is modeled according to a Markov chain with graphical representation shown in Fig. 1. Here $M$, $D$, $I$ denote match, delete and insert states, and the indices $B$ and $A$ refer to before the gene and after the gene. Further, the states GeneStart and GeneStop denote the start and stop of the gene, and $M_C$, $D_C$, $I_C$ denote the match, delete and insert codon states. The Begin state initializes the Markov chain, and the End state is used to model the random length of the sequences. In Section 2.1, we describe the transition probabilities between the hidden states in detail.

Each hidden state emits letters in the two sequences, and the number of letters emitted by each state can be seen in Table 1. Throughout we use the notation # for the presence of a letter and — for no letter being present. In the before and after gene states, single nucleotides are matched, deleted, or inserted, and in the remaining states nucleotide triplets are matched, deleted, or inserted. In Section 2.2, we describe the emission probabilities from each hidden state in detail.
FIG. 1. States and transitions of the pair HMM for prokaryotic gene structure prediction.

<table>
<thead>
<tr>
<th>Table 1. Letters Emitted from Each Hidden State$^a$</th>
<th>$M_B, M_A$</th>
<th>$D_B, D_A$</th>
<th>$I_B, I_A$</th>
<th>GeneStart, $M_C$, GeneStop</th>
<th>$D_C$</th>
<th>$I_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(###)</td>
<td>(#)</td>
<td>(#)</td>
<td>(###)</td>
<td>(###)</td>
<td>(#)</td>
<td>(#)</td>
</tr>
</tbody>
</table>

$^a$Here # denotes the presence of a letter (nucleotide) and - denotes the absence of a letter. The End and Begin state do not emit any letters.

2.1. Transition probabilities

In the three parts of the sequences corresponding to before the gene, the gene itself, and after the gene, we assume that the sequences have evolved according to the Thorne, Kishino, and Felsenstein (1991) model. In the TKF-model, each letter in an ancestral sequence develops independently of the other letters according to a birth and death process with birth rate $\lambda$ and death rate $\mu > \lambda$. This means that each ancestral letter is deleted after an exponentially distributed waiting time with mean $1/\mu$, and while the letter is present, it gives rise to new letters at the rate $\lambda$. New letters are placed immediately to the right of the letter giving birth and they are chosen from the stationary distribution of the substitution process. We assume that the birth and death rates are the same in the intergenic (before- and after-gene) regions of a sequence.

If an ancestral sequence has evolved to a present sequence during a time span $\tau$, the evolution can be summarized in terms of an alignment of some of the letters in the ancestral sequence with some of the letters in the present sequence (survival of these letters in the birth and death process), in terms of deletions (deaths) of some of the letters, in terms of insertions (births), and in terms of substitutions of the aligned letters. The TKF-model can be formulated as a Markov chain along the sequences with four states corresponding to match (survival with possible substitution), deletion of a single letter, insertion of a single letter, and an end state. The state-transition diagram of the TKF-model is depicted in Fig. 2.

As can be seen from Table 2, the transition probabilities of the TKF-model can be written as a product of at most three terms. The first term $b(\cdot, \cdot)$ represents the probability of having another birth $b(\cdot, #)$ (entering the I state) or having no more births $b(\cdot, -)$ (entering the M, D, or End state). The second term $\gamma$ represents the probability of having another letter in the ancestral sequence (entering the M or D states). Finally, the third term represents the probability of survival $s(\cdot)$ (entering the M state) or nonsurvival $s(-)$ (entering the D state) of a letter in the ancestral sequence during the evolution leading to the present sequence. To define these terms precisely, we introduce the short hand notation

$$\gamma = \frac{\lambda}{\mu} \quad \text{and} \quad \beta = \frac{1 - \exp(-(1 - \gamma)\mu\tau)}{1 - \gamma \exp(-(1 - \gamma)\mu\tau)}.$$ (2.1)

Then we have

$$b(#, #) = \gamma \beta, \quad b(#, -) = 1 - b(#, #),$$ (2.2)

$$b(-, #) = 1 - \frac{\beta}{1 - \exp(-\mu\tau)}, \quad b(-, -) = 1 - b(-, #),$$ (2.3)

$$s(#) = \exp(-\mu\tau), \quad s(-) = 1 - s(#).$$ (2.4)

Hein et al. (2003) give a careful introduction to the main probabilistic aspects of the TKF-model.
state-transition diagram of the TKF-model. When we use the TKF-model for the part of the DNA-sequence before the gene, the End state corresponds to the GeneStart state in Fig. 1. Similarly, when we use the TKF-model for the codon part of the sequences the End state corresponds to the GeneStop state in Fig. 1.

Table 2. Transition Probabilities between the Match, Delete, Insert and End States in the TKF-Model

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>D</th>
<th>I</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>b(#, -)γs(#)</td>
<td>b(#, -)γs(#)</td>
<td>b(#, #)</td>
<td>b(#, -)(1 - γ)</td>
</tr>
<tr>
<td>D</td>
<td>b(#)</td>
<td>b(#)</td>
<td>b(#, #)</td>
<td>b(#, -)(1 - γ)</td>
</tr>
<tr>
<td>I</td>
<td>b(#)</td>
<td>b(#)</td>
<td>b(#, #)</td>
<td>b(#, -)(1 - γ)</td>
</tr>
</tbody>
</table>

At the very left of the ancestral sequence is a birth process with rate \( λ \) so that the sequence will not eventually die out. This is achieved by letting the Begin state be a state with no emitted letters and where the transition probabilities are given by the first row of Table 2.

Note that the TKF-model has two parameters, \( γ \) and \( μτ \), and that the expected length \( EL \) of a sequence and the expected number of matches \( (EN_M|L) \), given that the ancestral sequence has length \( L \), are

\[
EL = \frac{γ}{1 - γ}, \quad (EN_M|L) = \exp(-μτ)L.
\]  

When we use the TKF-model for the part of the DNA sequence before the gene, the End state in Table 2 corresponds to the GeneStart state in Fig. 1, and the transition probabilities from the GeneStart states are given by the first row of Table 2 used for the codon part of the DNA sequences. Similarly, when we use the TKF-model for the codon part of the sequences, the End state in Table 2 corresponds to the GeneStop state in Fig. 1.

The transition probability of going from the hidden state \( x \) to the hidden state \( y \) in the Markov chain depicted in Fig. 1 is denoted \( p(x, y) \).

2.2. Emission probabilities

A state \( x \) emits letters in the positions where the symbol \( \# \) is present; see Table 1. We use the same emission probabilities in the before- and after-gene states. In the states \( D_B \) and \( D_A \), a nucleotide is emitted.
in sequence $S_1$, and the frequencies of the nucleotides
\[(\pi(A), \pi(G), \pi(C), \pi(T))\]
are assumed known. In the states $I_B$ and $I_A$, a nucleotide is emitted in sequence $S_2$ also from the distribution $\pi$. Finally in the states $M_B$ and $M_A$, a nucleotide $w_1$ is emitted in sequence $S_1$, and a nucleotide $w_2$ in sequence $S_2$. The distribution of this pair of nucleotides is
\[p_e(w_1, w_2) = \pi(w_1)f(w_2|w_1), \tag{2.6}\]
where $f(w_2|w_1)$ is the probability of a change from $w_1$ to $w_2$ within a time span $\tau_B$. We use an approximate form of the Hasegawa, Kishino, and Yano (1985) model for the substitution process corresponding to a small time span $\tau_B$. Thus, for $w_1 \neq w_2$ the probability of a change is
\[f(w_2|w_1) = \begin{cases} \tau_B\pi(w_2)/s_B & \text{for transition} \\ \kappa_B \tau_B\pi(w_2)/s_B & \text{for transversion}, \end{cases} \tag{2.7}\]
where $s_B$ is a scaling factor, and the probability of no change is
\[f(w_1|w_1) = 1 - \sum_{w_2 \neq w_1} f(w_2|w_1).\]

There are two parameters in the HKY-model, the time span between the sequences $\tau_B$ and the transition–transversion parameter $\kappa_B$. Usually, time is scaled such that it reflects the number of expected substitutions per site. In this case,
\[\tau_B = \sum_{w_i \neq w_j} p_e(w_i, w_j),\]
and so the scaling factor $s_B$ is given by
\[s_B = s(\kappa_B) \tag{2.8}\]
\[= 2 (\pi(A)\pi(G) + \pi(C)\pi(T)) + 2\kappa_B (\pi(A)\pi(C) + \pi(A)\pi(T) + \pi(G)\pi(C) + \pi(G)\pi(T)).\]

In the inside-gene states, sense codons are emitted. In the states $D_C$ and $I_C$, the frequency of the emitted nucleotide triplet is determined by the known distribution $\pi_C$. In the state $M_C$, a codon $w_1$ is emitted in sequence $S_1$, and a codon $w_2$ in $S_2$. We will use an approximate form of the Goldman and Yang (1994) model for the substitution process, where distances between amino acids are set to one. In this case, the Goldman and Yang model is given by the rate matrix
\[Q(w_1, w_2) = \begin{cases} 0 & \text{if } w_1 \text{ and } w_2 \text{ differ at more than one nucleotide} \\ \pi_C(w_2)/s_C & \text{for synonymous transition} \\ \kappa_C\pi_C(w_2)/s_C & \text{for synonymous transversion} \\ \omega_C\pi_C(w_2)/s_C & \text{for nonsynonymous transition} \\ \kappa_C\omega_C\pi_C(w_2)/s_C & \text{for nonsynonymous transversion}, \end{cases} \tag{2.9}\]
for $w_1 \neq w_2$, with corresponding substitution probabilities given by the matrix $\exp(Q\tau_C)$. We approximate this matrix by $I + Q\tau_C$ and add a term to take account of substitutions altering more than one nucleotide. Thus, we use the substitution probabilities
\[f(w_2|w_1) = \begin{cases} \theta_C\tau_C^2\pi_C(w_2)/s_C & \text{if } w_1 \text{ and } w_2 \text{ differ at more than one nucleotide} \\ \tau_C\pi_C(w_2)/s_C & \text{for synonymous transition} \\ \kappa_C\tau_C\pi_C(w_2)/s_C & \text{for synonymous transversion} \\ \omega_C\tau_C\pi_C(w_2)/s_C & \text{for nonsynonymous transition} \\ \rho_C\tau_C\pi_C(w_2)/s_C & \text{for nonsynonymous transversion}. \end{cases} \tag{2.10}\]
We have replaced $\kappa_c\omega_c$ by a free parameter $\rho_c$. The term with $\theta_c$ takes care of substitutions altering more than one nucleotide, and we scale this by $\tau_c^2$ to make such events less probable. There are five parameters in the approximate Goldman and Yang model, the time span $\tau_c$, and the four parameters $\kappa_c, \omega_c, \rho_c, \theta_c$ that distinguishes between synonymous transitions and transversions and nonsynonymous transitions and transversions and other types of substitutions. In this case, time is scaled such that it reflects the number of expected codon substitutions per codon site.

In the GeneStart state, the start codon ATG is emitted in sequences $S_1$ and $S_2$.

In the GeneStop state, stop codons are emitted in both sequences. As before, we assume that the distribution $\pi_S(w_1)$ of the stop codons TAA, TAG, and TGA is known. The conditional probabilities $f(w_2|w_1)$ are given in Table 3.

### 2.3. Parameter estimation

A summary of the 11 parameters of the model can be found in Table 4.

We estimate the parameters of the model by a modified version of the EM-algorithm. The EM-algorithm is a two-step maximization procedure. In the expectation step, mean values of a set of count statistics in the conditional distribution given the observed sequences and parameter values are calculated. In the maximization step, new parameter values are found by maximizing the full likelihood of the hidden states and the observed sequences with the counts replaced by their mean values.

Let $x_1, \ldots, x_n$ be the sequence of the hidden Markov chain generating the observed sequences $S_1$ and $S_2$ with $x_{n+1}$ being the End state of Fig. 1. Also, let $S[x_i] = (S_1[x_i], S_2[x_i])$ be the nucleotides emitted by $x_i$ in the sequence $x_1, \ldots, x_n$. The full likelihood of the sequences and the alignment is given by

$$L(\psi) = p(\text{Begin}, x_1) \prod_{i=1}^{n} p(x_i, x_{i+1}) p_e(S[x_i]|x_i),$$

where $\psi$ is the total set of parameters summarized in Table 4. When the nucleotide frequencies $\pi$ in the before- and after-gene states, codon frequencies $\pi_c$ in the inside-gene states, and stop codon frequencies $\pi_S$ in the stop-gene state are fixed, the full likelihood becomes, apart from a data dependent term,

$$L(\psi) = b(\#, \#)^N(\#, \#) b(\#, -)^N(\#, -) b(-, \#)^N(-, \#) b(-, -)^N(-, -) \times s(\#)^N(\#) s(-)^N(-) \gamma^N(1 - \gamma)^2 \prod_{w_1, w_2} f(w_2|w_1)^K(w_1, w_2)$$

$$\times b_c(\#, \#)^N_c(\#, \#) b_c(\#, -)^N_c(\#, -) b_c(-, \#)^N_c(-, \#) b_c(-, -)^N_c(-, -) \times s_c(\#)^N_c(\#) s_c(-)^N_c(-) \gamma_c^N_c(1 - \gamma_c) \prod_{w_1, w_2} f_c(w_2|w_1)^K_c(w_1, w_2)$$

$$\times f_{\text{GeneStop}}(w_2|w_1);$$

(2.11)

cf. Table 2 and (2.6). Here, we use index $c$ to indicate terms from the coding part of the alignment. The term $N(\#, \#)$ counts the number of times we make a transition to a before- or after-gene state with the

\begin{table}[h]
\centering
\caption{Conditional Probabilities $f(w_2|w_1)$ for the Nine Possible Emissions from the GeneStop State$^a$}
\begin{tabular}{ccc}
\hline
$w_2$ & TAA & TAG & TGA \\
\hline
$w_1$ & TAA & $\tau_c\pi_S(TAG)$ & $\tau_c\pi_S(TGA)$ \\
TAG & $\tau_c\pi_S(TAA)$ & TAA & $\theta_c^2\pi_S(TGA)$ \\
TGA & $\tau_c\pi_S(TAA)$ & $\theta_c^2\pi_S(TAG)$ & \\
\hline
\end{tabular}
\end{table}

$^a$Each row should sum to 1, giving the non-specified value in each row.
Table 4. Summary of the Parameters of the Pair Prokaryotic HMM

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>Start</th>
<th>Inside</th>
<th>Stop</th>
<th>After</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alignment</strong></td>
<td>γ_B, μ_Bτ_B</td>
<td>—</td>
<td>γ_C, μ_Cτ_C</td>
<td>—</td>
<td>γ_B, μ_Bτ_B</td>
<td>4</td>
</tr>
<tr>
<td><strong>Substitution</strong></td>
<td>τ_B, κ_B</td>
<td>—</td>
<td>τ_C, κ_C, ω_C, ρ_C, θ_C</td>
<td>τ_C, θ_C</td>
<td>τ_B, κ_B</td>
<td>7</td>
</tr>
</tbody>
</table>

The term \( b(#, #) \) in the transition probabilities, and all other counts are defined similarly. The term \( K(w_1, w_2) \) count, the number of times nucleotide \( w_1 \) is substituted by nucleotide \( w_2 \) in the before- or after-gene states, and \( K_C(w_1, w_2) \) counts the number of times codon \( w_1 \) is substituted by codon \( w_2 \) in the inside-gene states.

In the expectation step of the EM-algorithm, the mean values of all the counts have to be calculated, and in the maximization step, new parameter values are found by maximizing the full likelihood with the counts replaced by their mean values. However, the proposed pair HMM is rather complex, and the set of count statistics in (2.11) is large. For example, we need to count the expected number of substitutions between all 61 sense codons, a total of \( 61 \times 61 = 3721 \) counts. Therefore, we suggest to replace the maximization step in the EM-algorithm by a step where the new parameters are obtained from moment equations. In this modified EM-algorithm, the number of count statistics equals the number of parameters.

We now derive the moment equations. Consider the inside-gene states, and let \( N_{MC} \) be the number of matches, \( N_{DC} \) the number of deletions, and \( N_{IC} \) the number of insertions. From (2.5), we can write the two moment equations

\[
N_{MC} + N_{DC} = \gamma_C/(1 - \gamma_C) \quad \text{and} \quad N_{MC} + N_{IC} = \gamma_C/(1 - \gamma_C),
\]

where on the left sides we write the count and on the right sides we write the mean value of the count. We combine these into the equation

\[
\frac{N_{MC} + N_{DC} + N_{IC}}{2} = \frac{\gamma_C}{1 - \gamma_C}. \tag{2.12}
\]

Also, from (2.5), we have the moment equations

\[
N_{MC} = \exp(-\mu_C\tau_C)(N_{MC} + N_{DC}) \quad \text{and} \quad N_{MC} = \exp(-\mu_C\tau_C)(N_{MC} + N_{IC}),
\]

that are combined into

\[
N_{MC} = \exp(-\mu_C\tau_C) \left( N_{MC} + (N_{DC} + N_{IC})/2 \right). \tag{2.13}
\]

In the estimation step, we replace the count statistics in Equations (2.12) and (2.13) by their conditional mean values given the observed sequences. Solving the equations gives new parameter values of \( \gamma_C \) and \( \mu_C\tau_C \).

Similarly, the parameters of the TKF-model in the before- and after-gene states are estimated from the moment equations

\[
\frac{1}{2} \left( N_{MB} + \frac{N_{DB} + N_{IB}}{2} + N_{MA} + \frac{N_{DA} + N_{IA}}{2} \right) = \frac{\gamma_B}{1 - \gamma_B},
\]

\[
N_{MB} + N_{MA} = \exp(-\mu_B\tau_B) \left( N_{MB} + \frac{N_{DB} + N_{IB}}{2} + N_{MA} + \frac{N_{DA} + N_{IA}}{2} \right).
\]

Parameter estimation in the HKY-model is as follows. Recall that the model describes the substitution processes in the before- and after-gene states and that the parameters are \( \tau_B \) and \( \kappa_B \). Let \( N_{w_1/w_2} \) denote the
number of substitutions of \( w_1 \) by \( w_2 \) in the match before- or match after-gene states. From (2.7), we get, with \( w_1 \neq w_2 \), the moment equations

\[
N_{w_1w_2} = \begin{cases} \tau_B \pi(w_2) N_{w_1} / s_B & \text{for transition} \\ \kappa_B \tau_B \pi(w_2) N_{w_1} / s_B & \text{for transversion} \end{cases}, \quad \text{where } N_{w_1} = \sum_{w_2} N_{w_1w_2}.
\]

Adding all transition equalities and transversion equalities, we obtain

\[
N_{..} \tau_B = \frac{N_{AG}}{\pi(A)} + \frac{N_{GA}}{\pi(G)} + \frac{N_{CT}}{\pi(C)} + \frac{N_{TC}}{\pi(T)},
\]

(2.14)

\[
N_{..} \kappa_B \tau_B = \frac{N_{AC} + N_{AT} + N_{GC} + N_{GT}}{\pi(C)} + \frac{N_{CA} + N_{CG} + N_{TA} + N_{TG}}{\pi(A) + \pi(G)},
\]

(2.15)

where \( N_{..} = \sum_{w_1} N_{w_1} = N_{\mathcal{H}_6} + N_{\mathcal{H}_4} \) is the total number of matches in the before- and after-gene states. In the estimation step, we replace the count statistic in (2.14) and (2.15) by their conditional mean values given the observed sequences and obtain new values of \( \tau_B \) and \( \kappa_B \). Using the approximate form (2.7) of the HKY-model thus implies that parameter estimation requires three count statistics, namely, the conditional mean values of \( N_{\mathcal{H}_6} + N_{\mathcal{H}_4} \) and the conditional mean values of the right hand sides of (2.14) and (2.15).

In the appendix, we construct moment equations for parameter estimation in the original HKY-model.

In case of uniform frequencies \( \pi(A) = \pi(G) = \pi(C) = \pi(T) = 1/4 \), we get from (2.8), (2.14), and (2.15)

\[
s_B = \frac{1}{4} + \frac{\kappa_B}{2}, \quad N_{..} \tau_B = 4 N_{ts}, \quad N_{..} \kappa_B \tau_B = 2 N_{tv},
\]

where \( N_{ts} \) is the number of transitions and \( N_{tv} \) the number of transversions. Solving these equations, we obtain

\[
\tau_B = \frac{N_{ts}}{N_{..}}, \quad \kappa_B = \frac{N_{tv}}{2N_{ts}}.
\]

(2.16)

Thus, if the nucleotides are uniformly distributed, the time span \( \tau_B \) is estimated as the fraction of nucleotides undergoing changes, and the transversion–transition ratio \( \kappa_B \) is the fraction between the number of transversions and twice the number of transitions since there are twice as many possible transversions. If the nucleotides are not uniformly distributed, we obtain a weighted version of (2.16) as given by (2.8), (2.14), and (2.15).

The five parameters in the Goldman and Yang model are estimated in a similar way as for the HKY-model. The model describes the substitution process in the coding part of the sequences, and the parameters are \( \tau_C, \kappa_C, \omega_C, \rho_C, \theta_C \). Now let \( N_{w_1w_2} \) denote the number of codon substitutions of \( w_1 \) by \( w_2 \) in the match codon state. From (2.10), we get, with \( w_1 \neq w_2 \),

\[
N_{w_1w_2} = \begin{cases} \tau_C \pi(w_2) N_{w_1} / s_C & \text{for synonymous transition} \\ \kappa_C \tau_C \pi(w_2) N_{w_1} / s_C & \text{for synonymous transversion} \\ \omega_C \tau_C \pi(w_2) N_{w_1} / s_C & \text{for nonsynonymous transition} \\ \rho_C \tau_C \pi(w_2) N_{w_1} / s_C & \text{for nonsynonymous transversion} \\ \theta_C \tau_C^2 \pi(w_2) N_{w_1} / s_C & \text{otherwise} \end{cases}
\]

Adding, e.g., all synonymous transversion equalities, we obtain

\[
N_{..} \tau_C = \frac{\sum_{w_1} \sum_{w_2} \tau_C \pi(w_2) 1_{a,ts}(w_1, w_2)}{\sum_{w_1} \sum_{w_2} \pi(w_2) 1_{a,ts}(w_1, w_2)} = N_{a,ts},
\]

(2.17)

where \( 1_{a,ts}(w_1, w_2) \) is 1 if the change from \( w_1 \) to \( w_2 \) is a synonymous transition and 0 otherwise. Similar equalities can be obtained by adding equalities for synonymous transversions, nonsynonymous transitions,
nonsynonymous transversions, and other changes. Using the approximative form (2.10) of the Goldman and Yang model thus implies that parameter estimation requires six counts, namely, the conditional mean values of the number of codon matches $N_c = N_{bc}$ and conditional mean values of the right hand sides of the five equations similar to (2.17). In the appendix, we construct moment equations for parameter estimation in the original Goldman and Yang model (2.9).

We now describe how to calculate the count statistics. A subsequence of $S_j$ starting in $a$ and ending in $b$ is denoted $S_j[a:b]$, and if $a > b$ we interpret $S_j[a:b]$ as the empty set. Let $x$ be any state of the hidden Markov chain shown in Fig. 1, and let $K = (K_1, K_2)$ be numbers with $1 \leq K_i \leq L_i$. We then consider a recursion for the probability of a chain starting in the state $x$ generating the two sequences $S_1[K_1 : L_1]$ and $S_2[K_2 : L_2]$ from the states following $x$. Let us denote the latter probability by $P(K|x)$. The recursion is obtained by splitting the probability according to the value of the state following $x$ in the Markov chain. For a hidden state $y$, let $l(y) = (l_1(y), l_2(y))$ be the number of emitted nucleotides in the two sequences according to Table 1. For example, $l(M_B) = (1, 1)$, $l(M_B) = (1, 0)$, and $l(M_C) = (3, 3)$. Then the recursion is

$$P(K|x) = \sum_y p(x, y)p_e(K, l(y)|y)P(K + l(y)|y),$$  \hspace{1cm} (2.18)$$

where $p_e(K, l(y)|y)$ is the emission probability as described in Section 2.2 when emitting the nucleotides $S_1[K_1 : K_1 + l_1(y) - 1]$ in sequence $S_1$ and the nucleotides $S_2[K_2 : K_2 + l_2(y) - 1]$ in sequence $S_2$. In this notation the probability of the two sequences $S_1$ and $S_2$ is $P(1, 1|\text{Begin})$. Note that the sum in (2.18) always has four terms corresponding to the possible transitions in Table 2. The recursion is started at $(L_1 + 1, L_2 + 1) = L + 1$ and runs down to $(1, 1)$. The start of the recursion is given by

$$P(L + 1|x) = p(x, \text{End}),$$  \hspace{1cm} (2.19)$$

where End is the state shown in Fig. 1.

Let $x_1, \ldots, x_n$ be the sequence of the hidden Markov chain generating the observed sequences $S_1$ and $S_2$ with $x_{n+1}$ being the End state of Fig. 1. Further, let $S[x_i] = (S_1[x_i], S_2[x_i])$ be the nucleotides emitted by $x_i$ in the sequence $x_1, \ldots, x_n$. By a count statistic $N_A$, we mean a statistic of the form

$$N_A = \sum_{i=1}^{n} 1_A(x_i, S[x_i]),$$

where $A$ is some set. For example, $A$ could be defined such that

$$1_A(x, S[x]) = \begin{cases} 1 & \text{if } x = M_B \text{ and substituting } S_1[x] \text{ by } S_2[x] \text{ is a transition} \\ 0 & \text{otherwise}, \end{cases}$$

in which case $N_A$ is the number of transitions in the before-gene state. We want to be able to calculate the mean value of $N_A$ given the observed sequences $S_1$ and $S_2$. If a series of states ending in the state $x$ generated the sequences $S_1[1 : K_1 - 1]$ and $S_2[1 : K_2 - 1]$, we let $N_A(K|x)$ be the part of the count statistic $N_A$ that is due to the states following $x$. To calculate the conditional mean $EN_A(K|x)$ of $N_A(K|x)$ given the observed sequences, $K$, and $x$, we note that the conditional distribution of the first state $y$ following $x$ is

$$p(x, y)p_e(K, l(y)|y)P(K + l(y)|y)$$

$$P(K|x)$$

We therefore get the following recursion for $EN_A(K|x)$:

$$EN_A(K|x) = \sum_y (1_A(y, S[y]) + EN_A(K + l(y))) \frac{p(x, y)p_e(K, l(y)|y)P(K + l(y)|y)}{P(K|x)}.$$  \hspace{1cm} (2.20)$$

The start of the recursion is given by

$$EN_A(L + 1|x) = 0.$$
that the five substitution parameters fits the data. The parameters of the constrained Goldman and Yang model are estimated as follows. Recall \( A.tumefaciens \) and \( M.loti \) parts to the Markov chain depicted in Fig. 1. An intronic part can start in three possible phases, 0, 1, 2, 3.

### 2.4. Application to \( A.tumefaciens \) and \( M.loti \)

We applied the pair prokaryotic HMM to analyze two homologous sequences from \( Agrobacterium \) \( tumefaciens \) and \( Mesorhizobium \) \( loti \) of length 605 and 611 nucleotides, respectively. GenBank accession numbers are AE009042 and AP003011. We emphasize, as mentioned in the introduction, that the purpose is to make a close investigation of sequences found to be of interest by other means. The sequences code for parts of encoding an exodeoxyribonuclease small subunit. The modified EM-algorithm was constructed to have this property, but it is not ensured in the modified EM-algorithm. We also applied the algorithm with different starting values, and in each case the algorithm converged during a few iterations, and the result is summarized in Table 5. The first column in Table 5 shows the log probability of the sequences, and in this particular example the log likelihood increases after each iteration. The original EM-algorithm is constructed to have this property, but it is not ensured in the modified EM-algorithm. We also applied the algorithm with different starting values, and in each case the algorithm converged to the same parameters after a few iterations.

In Fig. 3, we indicate the gene structure prediction as obtained from the Viterbi algorithm with parameters inferred from the EM-algorithm.

We also investigated whether the constrained Goldman and Yang model given by (2.10) with \( \rho_C = \kappa_C \omega_C \) fits the data. The parameters of the constrained Goldman and Yang model are estimated as follows. Recall that the five substitution parameters \( \tau_C, \kappa_C, \omega_C, \rho_C, \theta_C \) in the coding part of the sequences are estimated from five equations of the type (2.17). In each iteration we therefore estimate the parameters of the constrained Goldman and Yang model by minimizing the sum of squares of differences between the left and right sides of these equations. Letting \( EN_{s,ts}, EN_{s,tv}, EN_{ns,ts}, EN_{ns,tv}, EN_{other} \) denote the counts on the right sides (with obvious notation), the estimates in each iteration minimize the sum of squares

\[
\left( \frac{EN_{H_s} \tau_C}{s_C} - EN_{s,ts} \right)^2 + \left( \frac{EN_{H_s} \kappa_C \tau_C}{s_C} - EN_{s,tv} \right)^2 + \left( \frac{EN_{H_s} \omega_C \tau_C}{s_C} - EN_{ns,ts} \right)^2
\]

\[
+ \left( \frac{EN_{H_s} \kappa_C \omega_C \tau_C}{s_C} - EN_{ns,tv} \right)^2 + \left( \frac{EN_{H_s} \theta_C \tau_C^2}{s_C} - EN_{other} \right)^2.
\]

The resulting parameter estimates and corresponding likelihood values are given in Table 5.

Carrying out a goodness-of-fit test of the pair prokaryotic HMM with the constrained Goldman and Yang model (2.10) with \( \rho_C = \kappa_C \omega_C \) under the prokaryotic HMM with the full Goldman and Yang model (2.10), we obtain a likelihood ratio test statistic equal to 6.3 on one degree of freedom. Using the \( \chi^2 \) approximation of the test statistic, the \( p \)-value is 1.2% and thus indicates that the full model fits significantly better than the constrained model.

### 3. PAIRWISE EUKARYOTIC GENE STRUCTURE PREDICTION AND ESTIMATION OF PARAMETERS

If the two homologous DNA sequences come from eukaryotic organisms, we have to introduce intronic parts to the Markov chain depicted in Fig. 1. An intronic part can start in three possible phases, 0, 1, 2.
FIG. 3. Part of the pairwise alignment of *A. tumefaciens* and *M. loti*. Light gray color corresponds to conserved positions, and nonconserved positions and gaps are shown in dark gray. The two black bars on top of the alignment indicate the start and stop of the gene.

depending on the codon reading frame. Further, we assume that the splice sites follow the GT–AG rule. According to this rule, an intronic part starts with the letters GT at a splice donor site and ends with the letters AG at a splice acceptor site. The graphical representation of the pair HMM for eukaryotic gene structure prediction is shown in Fig. 4.

In Fig. 4, the symbols $M_1$, $D_1$, $I_1$ denote match, delete, and insert intron states, and the IntronStart and IntronStop states denote the start and stop of the intron. The number of letters emitted by the match, delete, and insert intron states equals the numbers emitted from the match, delete, and insert before- and after-gene states. Similarly to the before- and after-gene states, the transition probabilities follow the TKF-model, and the emission probabilities are determined by the HKY-model with parameters specific for the intron states. The number of letters emitted by the IntronStart and IntronStop states can be seen in Table 6. In phase 0, the intronic part starts immediately after a sense codon. In phase 1, the first nucleotide in a codon is emitted in both sequences just before the donor splice site, and the codon is established by emitting two nucleotides in both sequences immediately after the acceptor site. Similarly, in phase 2, two nucleotides are emitted just before the donor splice site, and one nucleotide is emitted immediately after the acceptor site. Thus, the eukaryotic pair HMM maintains the reading frame across introns, but it does not prevent stop codons to occur across introns. To disallow stop codons, an extension of the three possible intron start states would be needed, where in phase 1 it is taken into account whether the nucleotide is a T or not, and in phase 2 whether the nucleotides are TA, TG, or not. Further extensions would be to allow gap triplets across introns and to keep track of codons across introns.

The probability of leaving the coding state from the match, delete, or insert states are given in the right column of Table 2, but having left the coding state, there are now two possible scenarios, namely, entering
an intron or ending the gene. Thus, the number of introns follow a geometric distribution with probability \( q \), say, of entering the intronic part. If we expect \( m \) intronic parts per gene, we fix \( q = m/(m+1) \).

Meyer and Durbin (2002) extend the model in Fig. 4 by allowing introns within untranslated regions of genes. They also allow introns which are present only in one of the genes. In an HMM, the duration time follows a geometric distribution. Pachter et al. (2002) relax this assumption in the exonic part of the model where they use a generalized HMM. Further extensions of the model include, e.g., sequencing errors and signals such as the TATA box in the promoter region of the gene and the Poly-A signal at the end of transcription; see Zhang (1998).

The EM-algorithm derived in Section 2.3 extends naturally to the case of two eukaryotic sequences with a common gene. The moment equations derived in Section 2.3 for the parameters in the before-, after-, and inside-gene states are similar, and the moment equations for the parameters in the intronic part of the sequences are similar to the equations in the before- and after-gene states. We applied the EM-algorithm to several homologous sequences from eukaryotic organisms, and in all cases the EM-algorithm converged to a maximum in a few iterations (results not shown).

4. TRIPLEWISE PROKARYOTIC GENE STRUCTURE PREDICTION

Now consider three homologous DNA sequences \( S_1, S_2, \) and \( S_3 \) of lengths \( L_1, L_2, \) and \( L_3 \) from prokaryotic organisms and suppose the sequences have one common gene. We use a hidden Markov model as in Fig. 1, but with the three states \{M, D, I\} replaced by a set of 15 states. This is because the TKF-model on a
3-star tree can be formulated as a hidden Markov model along the sequence with 15 states. In a 3-star tree, we have the observed sequences at the three leaves and an unobserved ancestral sequence at the interior node. The 15 states can be thought of as alignment columns with four entries compared to the HMM for pairwise prokaryotic gene structure prediction as shown in Fig. 1.

As in the case of two sequences, the transition probabilities are a product of at most three terms. The first term \( \prod b_j(\cdot, \cdot) \) represents the probability of having more births, the second term \( y \) represents the probability of having another letter in the ancestral sequence, and the third term represents the probability of survival of a new letter in the ancestral sequence. The precise formulation of the transition probability \( p(x, y) \) from state \( x = (x_0|x_1, x_2, x_3) \) to state \( y = (y_0|y_1, y_2, y_3) \) is given in Table 7, and we refer to Hein et al. (2003) for more details on the TKF-model for a 3-star tree.

As in the case of two sequences, at the very left of the ancestral sequence is a birth process with rate \( \lambda \), so that the sequence will not eventually die out. This is achieved by letting the Begin state be a state with no emitted letters and where the transition probabilities are given by the first row in Table 7 with \( x = (#|#|, #, #) \).

### 4.2. Emission probabilities

Recall that a hidden state \( x \) is an alignment column with four entries and that letters are emitted in those positions where the symbol # is present. First, consider the emission probabilities in the before- and after-gene states. Let the emitted letter be \( w = (w_j, j = 0, 1, 2, 3) \), where \( w_j \) is the empty set if \( x_j = - \).

### Table 7. Transition Probabilities in the 3-Star TKF-Model

<table>
<thead>
<tr>
<th>( y_0 = # )</th>
<th>( y_0 = - )</th>
<th>( y = \text{End} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_0 = # )</td>
<td>( \left{ \begin{array}{l} 3 \ j = 1 \end{array} b_j(x_j, -) \right} y \left{ \begin{array}{l} 3 \ j = 1 \end{array} s_j(y_j) \right} \right} \left{ \begin{array}{l} 3 \ j = 1 \end{array} b_j(x_j, y_j) \right} \left{ \begin{array}{l} 3 \ j = 1 \end{array} b_j(x_j, -) \right} (1 - y) \right}</td>
<td>( x_0 = - )</td>
</tr>
</tbody>
</table>

\(^a\)The terms \( b_j(\cdot, \cdot) \) and \( s_j(\cdot) \) are defined in (2.2)–(2.4) with \( \tau \) replaced by \( \tau_j, j = 1, 2, 3 \). Transitions from a state with \( x_0 = - \) to a state with \( y_0 = - \) is possible only if \( y_j = - \) when \( x_j = - \).
Following (2.6), the emission probabilities in the before- and after-gene states are given by

\[
p_e^0(w|x) = \begin{cases} 
\pi(w_0) \prod_{j \geq 1, x_j = \#} f_j(w_j|w_0) & \text{if } x_0 = \#

\prod_{j \geq 1, x_j = \#} \pi(w_j) & \text{if } x_0 = -. 
\end{cases}
\]  

(4.1)

Here, \( f_j(w_j|w_0) \) is given by (2.7) with the intergenic evolutionary distance \( \tau_B \) and the transition–transversion parameter \( \kappa_B \) replaced by branch-specific parameters \( \tau_{B,j}, j = 1, 2, 3, \) and \( \kappa_{B,j}, j = 1, 2, 3. \)

The marginal probability of \((w_1, w_2, w_3)\) given the state \(x\) is obtained by summing over \(w_0\) in the previous expression

\[
p_e((w_1, w_2, w_3)|x) = \begin{cases} 
\sum_{w_0} \pi(w_0) \prod_{j \geq 1, x_j = \#} f_j(w_j|w_0) & \text{if } x_0 = \#

\prod_{j \geq 1, x_j = \#} \pi(w_j) & \text{if } x_0 = -. 
\end{cases}
\]  

(4.2)

In particular, if \((w_1, w_2, w_3) = (w_1, -, -)\) and \(x_0 = \#\), we get

\[
p_e((w_1, -, -)|(\#|\#, -, -)) = \sum_{w_0} \pi(w_0) f(w_1|w_0) = \pi(w_1),
\]

since \(\pi\) is the stationary distribution. We are now in a position to find the conditional distribution of a letter in the unobserved ancestral sequence given the letters at the three leaves and the hidden state. This probability is given by

\[
p_e^0(w_0|(w_1, w_2, w_3), x) = \frac{p_e^0(w|x)}{p_e((w_1, w_2, w_3)|x)},
\]  

(4.3)

where the probabilities on the right side are given by (4.1) and (4.2).

The emission probabilities in the inside-gene states are defined as in (4.1) with \(\pi\) replaced by \(\pi_C\) and with \(f_j(\cdot, \cdot)\) determined by (2.10). Here, the evolutionary distance \(\tau_C\) and the parameters \(\kappa_C, \omega_C, \rho_C, \theta_C\) are replaced by branch-specific parameters.

The GeneStart state emits the start codon ATG in all three observed sequences and in the ancestral sequence.

In the GeneStop state, stop codons are emitted in all four sequences, and the emission probabilities are determined by Table 3 and

\[
p_e^0(w|x) = \pi_B(w_0) \prod_{j=1}^3 f_j(w_j|w_0), \quad w_j \in \{\text{TAA}, \text{TAG}, \text{TGA}\}, \quad j = 0, 1, 2, 3.
\]

4.3. Parameter estimation

Recall the 11 parameters of the pairwise prokaryotic HMM summarized in Table 4. In the 3-star HMM, we let the parameters of the TKF-model \(\gamma_B, \mu_B, \gamma_C, \mu_C\) be common parameters on all lineages. For the remaining parameters, we consider the full model with \(\tau_B, \kappa_B, \tau_C, \kappa_C, \omega_C, \rho_C, \theta_C\) being branch specific. Again we use a modified EM-algorithm based on moment equations to estimate the parameters in the 3-star HMM.

Consider the inside-gene states, and let \(N_{\kappa_C}\) be the number of states having the symbol ### in the ancestral sequence and \(N_{\text{FMC}}\) be the number of full matches (###|###, ###|###). From (2.5), we can write the moment equations

\[
N_{\kappa_C} = \gamma_C/(1 - \gamma_C), \quad N_{\text{FMC}} = \exp \left( -\mu_C(\tau_{C,1} + \tau_{C,2} + \tau_{C,3}) \right) N_{\kappa_C}.
\]  

(4.4)
In the estimation step, we replace the count statistic in (4.4) by their conditional mean values given the observed sequences.

Because of the silent states \( Q_B = (\#, -, -, -) \), \( Q_A = (\#|−, −, −) \), and \( Q_C = (\###|−−, −−−, −−−) \), the start of the recursion (2.19) and the recursion (2.18) become more complicated. With \( (L_1, L_2, L_3) = L \), the start of the recursion now becomes

\[
P(L + 1|Q_A) = \frac{p(Q_A, \text{End})}{1 - p(Q_A, Q_A)},
\]

and for \( x \neq Q_A \),

\[
P(L + 1|x) = p(x, \text{End}) + p(x, Q_A)P(L + 1|Q_A),
\]

where End is the state shown in Fig. 5. The recursion is given first by finding the marginal probability of the sequences \( S[K : L + 1] \) given that the initial state \( Q \) is one of the silent states \( Q_B, Q_A, Q_C \).

\[
P(K|Q) = \frac{1}{1 - p(Q, Q)} \sum_{y \neq Q} p(Q, y)p_e(K, l(y)|y)P(K + l(y)|y),
\]

and second by finding the marginal probability for the nonsilent states as in (2.18).

Parameter estimation of the substitution probabilities is complicated by the fact that the letters of the ancestral sequence are unknown. Calculating the conditional mean given the observed sequences therefore involves an extra step where the mean over the ancestral letter is calculated. This is done via (4.3) and amounts to replacing the indicator function in (2.20) by

\[
\sum_{w_0} l_A(y, w_0, S[y])p^0_e(w_0|S[y], y).
\]

4.4. Application to A.tumefaciens, M.loti and S.meliloti

We applied the 3-star prokaryotic HMM to analyze homologous sequences from *Agrobacterium tumefaciens*, *Mesorhizobium loti*, and *Sinorhizobium meliloti*. The first two sequences are described in Section 2.4, and the last has GenBank accession number AP003011 and length 601 nucleotides. We used the parameters from the pairwise comparisons of the sequences as starting values for the 3-star EM-algorithm. With these starting values, the EM-algorithm converged after a few iterations. In Table 8, we show the final parameter estimates, and in Fig. 6 we indicate a part of the gene structure prediction as obtained from the Viterbi algorithm.

In the 3-star model, one may wish to consider several different constrained models. For example, one may expect the transition–transversion parameters \( \kappa_B \) and \( \kappa_C \) to be the same in all branches, and perhaps even the same in the intergenic and coding parts. The synonymous–nonsynonymous ratio \( \omega_C \) is of interest on its own since a value of \( \omega_C \) larger than one indicates positive selection; cf. Nielsen and Yang (1998). In this particular data example, this is surely not the case. Further, it is natural to assume the evolutionary distances to scale linearly in the intergenic and coding regions such that

\[
(\tau_{1,B}, \tau_{2,B}, \tau_{3,B}) = \xi(\tau_{1,C}, \tau_{2,C}, \tau_{3,C}), \quad \xi > 0.
\]

In Section 2.4, we discussed how to fit constrained models by minimizing a certain sum of squares.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>( \tau_B )</th>
<th>( \kappa_B )</th>
<th>( \tau_C )</th>
<th>( \kappa_C )</th>
<th>( \omega_C )</th>
<th>( \rho_C )</th>
<th>( \theta_C )</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A.tumefaciens</em></td>
<td>0.188</td>
<td>0.629</td>
<td>0.315</td>
<td>0.257</td>
<td>0.010</td>
<td>0.076</td>
<td>0.063</td>
</tr>
<tr>
<td><em>M.loti</em></td>
<td>0.331</td>
<td>0.855</td>
<td>0.367</td>
<td>0.765</td>
<td>0.397</td>
<td>0.632</td>
<td>0.386</td>
</tr>
<tr>
<td><em>S.meliloti</em></td>
<td>0.271</td>
<td>0.807</td>
<td>0.371</td>
<td>0.266</td>
<td>0.002</td>
<td>0.001</td>
<td>0.026</td>
</tr>
</tbody>
</table>

\( ^a \gamma_B = 0.994, \mu_B = 0.064, \gamma_C = 0.988, \mu_C = 0.002, l = -2128.5. \)
 CHARACTERIZATION OF HOMOLOGOUS DNA SEQUENCES

We fitted the constrained model with $\kappa_B$ being the same in all branches. The fitted value of $\kappa_B$ is 0.788, and the remaining parameter values changed only slightly compared to the full model. The log likelihood is $-2128.7$, and so we obtain a likelihood ratio test statistic equal to 0.4 on two degrees of freedom. Using the $\chi^2(2)$ approximation of the test statistic, the $p$-value is 82% and thus supports the expectation that the constrained model is sufficiently flexible compared to the full model.

5. DISCUSSION

As demonstrated by Hein et al. (2003), the TKF-model can be written as a hidden Markov chain along any number of sequences related by a phylogenetic tree. Therefore, the hidden Markov model introduced in this paper can in principle be extended to any number of species. In this paper, we have combined the TKF-model with the Hasegawa, Kishino, and Yano (1985) substitution process in the before- and after-gene states and the Goldman and Yang (1994) substitution process in the inside-gene states, but other substitution processes also apply.

The EM-algorithm for a pair HMM with $S$ states and $T$ transitions and sequences of length $L_1 < L_2$ requires time of the order $O(SL_1)$ and memory of the order $O(TL_1L_2)$. For a triple HMM and sequences of length $L_1 < L_2 < L_3$, the time and memory requirements are of the order $O(SL_1L_2)$ and $O(TL_1L_2L_3)$. For two sequences, Meyer and Durbin (2002) developed the stepping stone algorithm, where subsequences of strong similarity are used as fixed points for the alignment. Pachter et al. (2002) use the global sequence alignment system GLASS described by Batzoglou et al. (2000) as an anchoring method to reduce the alignment space. Similar algorithms can be formulated for multiple sequences and are needed if alignment and gene structure prediction are carried out simultaneously. An alternative approach is to search the alignment space using simulation procedures as discussed by Holmes and Bruno (2001) and Jensen and Hein (2004).

In this paper, we have described our estimation procedure as a natural approximation to the EM-algorithm. It seems plausible that general convergence results can be obtained by recasting the method within the framework of stochastic Robbins–Monro approximations (Robbins and Monro, 1951).

APPENDIX

In this Appendix, we construct moment equations for parameter estimation in the HKY and Goldman and Yang models.

The probability for a pair of nucleotides in the HKY-model is determined by the rate matrix

$$Q(w_1, w_2) = \begin{cases} \tau_B \pi(w_2)/s_B & \text{for transition} \\ \kappa_B \tau_B \pi(w_2)/s_B & \text{for transversion} \end{cases}$$
for \( w_1 \neq w_2 \), with corresponding substitution probabilities given by the matrix \( \exp(Q \tau_B) \). Let \( N_w \) denote the number of times \( w \) occur in sequence \( S_1 \) in the before or after match states. Further let \( N_{w_1 w_2} \) denote the number of times \( w_1 \) in sequence \( S_1 \) is substituted with \( w_2 \) in sequence \( S_2 \) in the before or after match states. We may then estimate \( \tau_B \) and \( \kappa_B \) from the two moment equations

\[
N_{AG} + N_{GA} + N_{CT} + N_{TC} = N_A p_{AG} (\tau_B, \kappa_B) + N_G p_{GA} (\tau_B, \kappa_B) + N_C p_{CT} (\tau_B, \kappa_B) + N_T p_{TC} (\tau_B, \kappa_B), \tag{A.1}
\]

and

\[
N_{AC} + N_{AT} + N_{CG} + N_{GT} + N_{CA} + N_{CG} + N_{TA} + N_{TG} = N_A (p_{AC} (\tau_B, \kappa_B) + p_{AT} (\tau_B, \kappa_B)) + N_G (p_{GC} (\tau_B, \kappa_B) + p_{GT} (\tau_B, \kappa_B)) + N_C (p_{CA} (\tau_B, \kappa_B) + p_{CG} (\tau_B, \kappa_B)) + N_T (p_{TA} (\tau_B, \kappa_B) + p_{TG} (\tau_B, \kappa_B)), \tag{A.2}
\]

where \( p_{w_1 w_2} (\tau_B, \kappa_B) \) is the \((w_1, w_2)\)th entry in \( \exp(Q \tau_B) \). The equations would have to be solved numerically and require six counts, namely, \( N_A, N_G, N_C, N_T \), and the left sides of (A.1) and (A.2).

In the Goldman and Yang model (2.9) with \( \kappa_C \omega_C \) replaced by the free parameter \( \rho_C \), we get with a similar notation,

\[
\sum_{w_1, w_2} N_{w_1 w_2} 1_{s, ts}(w_1, w_2) = \sum_{w_1, w_2} N_{w_1 w_2} p_{w_1 w_2} (\tau_C, \kappa_C, \omega_C, \rho_C) 1_{s, ts}(w_1, w_2),
\]

where \( 1_{s, ts}(w_1, w_2) \) is 1 if the change from \( w_1 \) to \( w_2 \) is a synonymous transition and 0 otherwise. Similarly, three other equations with \( 1_{s, ts} \) replaced by \( 1_{s, tv} \) (synonymous transversions), \( 1_{ns, ts} \) (nonsynonymous transitions), and \( 1_{ns, tv} \) (nonsynonymous transversions) are obtained. These four equations should be solved numerically and require 65 counts, namely, the 61 sense codon counts \( N_w \) and the four counts of the left sides.

ACKNOWLEDGMENTS

We thank Jacob S. Pedersen for sharing his knowledge on gene finding and Leif Schauser for providing the data. We would also like to thank the referee for a very valuable and useful report. This work was supported by the Aarhus University Research Foundation and grant 21-00-0283 from the Danish Natural Science Research Council.

REFERENCES


Address correspondence to:
Asger Hobolth
Bioinformatics Research Center
University of Aarhus
Hoegh-Guldbergs Gade 10
Building 090
DK-8000 Aarhus C
Denmark

E-mail: asger@birc.au.dk