Applications of hidden Markov models for comparative gene structure prediction

Asger Hobolth[1] and Jens Ledet Jensen[2]

[1] *Bioinformatics Research Centre, University of Aarhus*
[2] *Department of Theoretical Statistics, University of Aarhus*

ABSTRACT. Identifying 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. The model integrates structure prediction methods for one sequence, statistical multiple alignment methods and phylogenetic information and is applied to 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 provides 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 part 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 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

This work was supported by MaPhySto – A Network in Mathematical Physics and Stochastics, funded by The Danish National Research Foundation.

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 maximisation (EM) algorithm and moment equations.

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 corresponds 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 non-synonymous and synonymous codon substitutions. The start, stop, donor and acceptor site positions are modelled in terms of simple functional signals.

Pedersen and Hein (2003) also predict gene structure in multiple related sequences, but their HMM assumes that alignment of the sequences have 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 next 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 finishes 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 ith 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 modelled according to a Markov chain with graphical representation shown in Figure 1. Here M, D, I denote match, delete and insert states, and the indices B and A refers to before the gene and after the gene. Further the states GeneStart and GeneStop denote the start and stop of the gene, and $\mathtt{M}_\mathtt{C}, \mathtt{D}_\mathtt{C}, \mathtt{I}_\mathtt{C}$ 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.

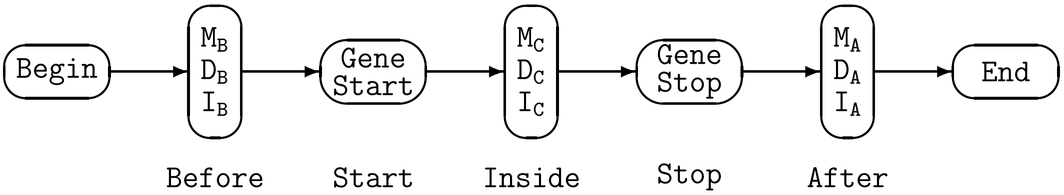
Begin
$\mathtt{M}_\mathtt{B}$ $\mathtt{D}_\mathtt{B}$ $\mathtt{I}_\mathtt{B}$
Gene Start
$\mathtt{M}_\mathtt{C}$ $\mathtt{D}_\mathtt{C}$ $\mathtt{I}_\mathtt{C}$
Gene Stop
$\mathtt{M}_\mathtt{A}$ $\mathtt{D}_\mathtt{A}$ $\mathtt{I}_\mathtt{A}$
End
Before
Start
Inside
Stop
After

Figure 1: States and transitions of the pair HMM for prokaryotic gene structure prediction.

Each hidden state emits letters in the two sequences, and the number of letters emitted by each state can be seen in Table 1. We use throughout 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.

$\mathtt{M}_\mathtt{B}$, $\mathtt{M}_\mathtt{A}$	$\mathtt{D}_\mathtt{B}$, $\mathtt{D}_\mathtt{A}$	$\mathtt{I}_\mathtt{B}$, $\mathtt{I}_\mathtt{A}$	GeneStart, $\mathtt{M}_\mathtt{C}$, GeneStop	$\mathtt{D}_\mathtt{C}$	$\mathtt{I}_\mathtt{C}$
$\begin{pmatrix}\#\\\#\end{pmatrix}$	$\begin{pmatrix}\#\\-\end{pmatrix}$	$\begin{pmatrix}-\\\#\end{pmatrix}$	$\begin{pmatrix}\#\#\#\\\#\#\#\end{pmatrix}$	$\begin{pmatrix}\#\#\#\\---\end{pmatrix}$	$\begin{pmatrix}---\\\#\#\#\end{pmatrix}$

Table 1: Letters emitted from each hidden state. 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 λ 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 λ. New letters are placed immediately to the right of the letter giving birth and is 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 τ 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 three states corresponding to match (survival with possible substitution), deletion of a single letter, insertion of a single letter, and and an end state.

The transition probabilities can be written as a product of at most three terms. The first terms $b(\cdot,\cdot)$ represents the probability of having another birth $b(\cdot,\#)$ or having no more births $b(\cdot,-)$. The second term γ represents the probability of having another letter in the ancestral sequence. Finally, the third term $s(\cdot)$ represents the probability of survival of a new letter in the ancestral sequence. The precise definition of these terms are

$$\beta = \frac{1-\exp(-(1-\gamma)\mu\tau)}{1-\gamma\exp(-(1-\gamma)\mu\tau)}, \quad \gamma = \lambda/\mu, \tag{2.1}$$

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

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

$$s(\#) = \exp(-\mu\tau), \quad s(-) = 1 - s(\#), \tag{2.4}$$

The transition probabilities in the TKF-model can be seen in Table 2.

	M	D	I	End
M	$b(\#,-)\gamma s(\#)$	$b(\#,-)\gamma s(-)$	$b(\#,\#)$	$b(\#,-)(1-\gamma)$
D	$b(-,-)\gamma s(\#)$	$b(-,-)\gamma s(-)$	$b(-,\#)$	$b(-,-)(1-\gamma)$
I	$b(\#,-)\gamma s(\#)$	$b(\#,-)\gamma s(-)$	$b(\#,\#)$	$b(\#,-)(1-\gamma)$

Table 2: Transition probabilities between the match, delete, insert and end states in the TKF-model.

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 $\mu\tau$, and that the expected length EL of a sequence and the expected number of matches $(EN_{\mathtt{M}}|L)$ given that the sequence has length L are

$$EL = \gamma/(1-\gamma), \quad (EN_{\mathtt{M}}|L) = \exp(-\mu\tau)L. \tag{2.5}$$

Hein *et al.* (2003) give a careful introduction to the main probabilistic aspects of the TKF-model.

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 Figure 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 Figure 1.

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

2.2 Emission probabilities

A state x emits letters in those positions where the symbol # is present. We use the same emission probabilities in the before and after gene states. In the states $\mathtt{D_B}$ and $\mathtt{D_A}$ a nucleotide is emitted in sequence S_1, and the frequencies of the nucleotides

$$(\pi(\mathtt{A}), \pi(\mathtt{G}), \pi(\mathtt{C}), \pi(\mathtt{T}))$$

are assumed known. In the states $\mathtt{I_B}$ and $\mathtt{I_A}$ a nucleotide is emitted in sequence S_2 also from the distribution π. Finally in the states $\mathtt{M_B}$ and $\mathtt{M_A}$ 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 $\tau_{\mathtt{B}}$. We use an approximative form of the Hasegawa, Kishino and Yano (1985) model for the substitution process corresponding to a small time span $\tau_{\mathtt{B}}$. Thus for $w_1 \neq w_2$ the probability of a change is

$$f(w_2|w_1) = \begin{cases} \tau_{\mathtt{B}} \pi(w_2)/s_{\mathtt{B}} & \text{for transition} \\ \kappa_{\mathtt{B}} \tau_{\mathtt{B}} \pi(w_2)/s_{\mathtt{B}} & \text{for transversion,} \end{cases} \tag{2.7}$$

where $s_{\mathtt{B}}$ 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 $\tau_{\mathtt{B}}$ and the transition-transversion parameter $\kappa_{\mathtt{B}}$. Usually time is scaled such that it reflects the number of expected substitutions per site. In this case

$$\tau_{\mathtt{B}} = \sum_{w_i \neq w_j} p_e(w_i, w_j),$$

and so the scaling factor $s_{\mathtt{B}}$ is given by

$$\begin{aligned} s_{\mathtt{B}} &= s(\kappa_{\mathtt{B}}) \qquad (2.8) \\ &= 2\Big(\pi(\mathtt{A})\pi(\mathtt{G}) + \pi(\mathtt{C})\pi(\mathtt{T})\Big) + 2\kappa_{\mathtt{B}}\Big(\pi(\mathtt{A})\pi(\mathtt{C}) + \pi(\mathtt{A})\pi(\mathtt{T}) + \pi(\mathtt{G})\pi(\mathtt{C}) + \pi(\mathtt{G})\pi(\mathtt{T})\Big). \end{aligned}$$

In the inside gene states sense codons are emitted. In the states $\mathtt{D}_\mathtt{C}$ and $\mathtt{I}_\mathtt{C}$ the frequency of the emitted nucleotide triplet is determined by the known distribution $\pi_\mathtt{C}$. In the state $\mathtt{M}_\mathtt{C}$ 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. For the simplified case of identical distances between amino-acids the probability for a pair of emitted codons in the Goldman and Yang model is given by the rate matrix

$$Q(w_1, w_2) = \begin{cases} \pi_\mathtt{C}(w_2)/s_\mathtt{C} & \text{for synonymous transition} \\ \kappa_\mathtt{C}\pi_\mathtt{C}(w_2)/s_\mathtt{C} & \text{for synonymous transversion} \\ \omega_\mathtt{C}\pi_\mathtt{C}(w_2)/s_\mathtt{C} & \text{for nonsynonymous transition} \\ \kappa_\mathtt{C}\omega_\mathtt{C}\pi_\mathtt{C}(w_2)/s_\mathtt{C} & \text{for nonsynonymous transversion} \\ 0 & \text{otherwise,} \end{cases} \tag{2.9}$$

for $w_1 \neq w_2$, with corresponding substitution probabilities given by the matrix $\exp(Q\tau_\mathtt{C})$. We approximate this matrix by $I + Q\tau_\mathtt{C}$ and add a term to take account of substitutions altering more than one codon. Thus we use the substitution probabilities

$$f(w_2|w_1) = \begin{cases} \tau_\mathtt{C}\pi_\mathtt{C}(w_2)/s_\mathtt{C} & \text{for synonymous transition} \\ \kappa_\mathtt{C}\tau_\mathtt{C}\pi_\mathtt{C}(w_2)/s_\mathtt{C} & \text{for synonymous transversion} \\ \omega_\mathtt{C}\tau_\mathtt{C}\pi_\mathtt{C}(w_2)/s_\mathtt{C} & \text{for nonsynonymous transition} \\ \rho_\mathtt{C}\tau_\mathtt{C}\pi_\mathtt{C}(w_2)/s_\mathtt{C} & \text{for nonsynonymous transversion} \\ \theta_\mathtt{C}\tau_\mathtt{C}^2\pi_\mathtt{C}(w_2)/s_\mathtt{C} & \text{otherwise.} \end{cases} \tag{2.10}$$

We have replaced $\kappa_\mathtt{C}\omega_\mathtt{C}$ by a free parameter $\rho_\mathtt{C}$. The term with $\theta_\mathtt{C}$ takes care of substitutions altering more than one codon and we scale this by $\tau_\mathtt{C}^2$ to make such events less probable. There are five parameters in the approximate Goldman and Yang model, the time span $\tau_\mathtt{C}$, and the four parameters $\kappa_\mathtt{C}, \omega_\mathtt{C}, \rho_\mathtt{C}, \theta_\mathtt{C}$ 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 sequence S_1 and S_2.

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

		w_2		
		TAA	TAG	TGA
	TAA	$\cdot$	$\tau_\mathtt{C}\pi_\mathtt{S}(\mathtt{TAG})$	$\tau_\mathtt{C}\pi_\mathtt{S}(\mathtt{TGA})$
w_1	TAG	$\tau_\mathtt{C}\pi_\mathtt{S}(\mathtt{TAA})$	$\cdot$	$\theta_\mathtt{C}\tau_\mathtt{C}^2\pi_\mathtt{S}(\mathtt{TGA})$
	TGA	$\tau_\mathtt{C}\pi_\mathtt{S}(\mathtt{TAA})$	$\theta_\mathtt{C}\tau_\mathtt{C}^2\pi_\mathtt{S}(\mathtt{TAG})$	$\cdot$

Table 3: Conditional probabilities $f(w_2|w_1)$ for the nine possible emissions from the `GeneStop` state. Each row should sum to 1, giving the non-specified value in each row.

2.3 Parameter estimation

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

	Before	Start	Inside	Stop	After	total
Alignment	$\gamma_{\mathtt{B}}$, $\mu_{\mathtt{B}}\tau_{\mathtt{B}}$	-	$\gamma_{\mathtt{C}}$, $\mu_{\mathtt{C}}\tau_{\mathtt{C}}$	-	$\gamma_{\mathtt{B}}$, $\mu_{\mathtt{B}}\tau_{\mathtt{B}}$	4
Substitution	$\tau_{\mathtt{B}}$, $\kappa_{\mathtt{B}}$	-	$\tau_{\mathtt{C}}$, $\kappa_{\mathtt{C}}$, $\omega_{\mathtt{C}}$, $\rho_{\mathtt{C}}$, $\theta_{\mathtt{C}}$	$\tau_{\mathtt{C}}$, $\theta_{\mathtt{C}}$	$\tau_{\mathtt{B}}$, $\kappa_{\mathtt{B}}$	7

Table 4: Summary of the parameters of the pair prokaryotic HMM.

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 distribution of the hidden states and the observed sequences with the counts replaced by their mean values. The proposed pair HMM is rather complex, and the set of count statistics is large. Therefore we suggest replacing the maximization step by a moment step such that the parameters are estimated from moment equations. In this modified EM-algorithm the number of count statistics equals the number of parameters.

Consider the inside gene states and let $N_{\mathtt{M_C}}$ be the number of matches, $N_{\mathtt{D_C}}$ the number of deletions, and $N_{\mathtt{I_C}}$ the number of insertions. From (2.5) we can write the two moment equations

$$N_{\mathtt{M_C}} + N_{\mathtt{D_C}} = \gamma_{\mathtt{C}}/(1-\gamma_{\mathtt{C}}) \quad \text{and} \quad N_{\mathtt{M_C}} + N_{\mathtt{I_C}} = \gamma_{\mathtt{C}}/(1-\gamma_{\mathtt{C}}).$$

We combine these into the equation

$$N_{\mathtt{M_C}} + (N_{\mathtt{D_C}} + N_{\mathtt{I_C}})/2 = \gamma_{\mathtt{C}}/(1-\gamma_{\mathtt{C}}). \tag{2.11}$$

Also from (2.5) we have the moment equations

$$N_{\mathtt{M_C}} = \exp(-\mu_{\mathtt{C}}\tau_{\mathtt{C}})(N_{\mathtt{M_C}} + N_{\mathtt{D_C}}) \quad \text{and} \quad N_{\mathtt{M_C}} = \exp(-\mu_{\mathtt{C}}\tau_{\mathtt{C}})(N_{\mathtt{M_C}} + N_{\mathtt{I_C}}),$$

that are combined into

$$N_{\mathtt{M_C}} = \exp(-\mu_{\mathtt{C}}\tau_{\mathtt{C}})\Big(N_{\mathtt{M_C}} + (N_{\mathtt{D_C}} + N_{\mathtt{I_C}})/2\Big). \tag{2.12}$$

In the estimation step we replace the count statistics in (2.11) and (2.12) by their conditional mean values given the observed sequences.

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

$$\frac{1}{2}\Big(N_{\mathtt{M_B}} + \frac{N_{\mathtt{D_B}} + N_{\mathtt{I_B}}}{2} + N_{\mathtt{M_A}} + \frac{N_{\mathtt{D_A}} + N_{\mathtt{I_A}}}{2}\Big) = \frac{\gamma_{\mathtt{B}}}{1-\gamma_{\mathtt{B}}}$$
$$(N_{\mathtt{M_B}} + N_{\mathtt{M_A}}) = \exp(-\mu_{\mathtt{B}}\tau_{\mathtt{B}})\Big(N_{\mathtt{M_B}} + \frac{N_{\mathtt{D_B}} + N_{\mathtt{I_B}}}{2} + N_{\mathtt{M_A}} + \frac{N_{\mathtt{D_A}} + N_{\mathtt{I_A}}}{2}\Big).$$

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 $\tau_{\mathtt{B}}$ and $\kappa_{\mathtt{B}}$. 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

$$N_{w_1 w_2} = \begin{cases} \tau_{\mathtt{B}} \pi(w_2) N_{w_1 \cdot}/s_{\mathtt{B}} & \text{for transition} \\ \kappa_{\mathtt{B}} \tau_{\mathtt{B}} \pi(w_2) N_{w_1 \cdot}/s_{\mathtt{B}} & \text{for transversion,} \end{cases} \qquad \text{where } N_{w_1 \cdot} = \sum_{w_2} N_{w_1 w_2}.$$

Adding all transition equalities and transversion equalities we obtain

$$N_{\cdot\cdot} \frac{\tau_{\mathtt{B}}}{s_{\mathtt{B}}} = \frac{N_{\mathtt{AG}}}{\pi(\mathtt{A})} + \frac{N_{\mathtt{GA}}}{\pi(\mathtt{G})} + \frac{N_{\mathtt{CT}}}{\pi(\mathtt{C})} + \frac{N_{\mathtt{TC}}}{\pi(\mathtt{T})} \tag{2.13}$$

$$N_{\cdot\cdot} \kappa_{\mathtt{B}} \frac{\tau_{\mathtt{B}}}{s_{\mathtt{B}}} = \frac{N_{\mathtt{AC}} + N_{\mathtt{AT}} + N_{\mathtt{GC}} + N_{\mathtt{GT}}}{\pi(\mathtt{C}) + \pi(\mathtt{T})} + \frac{N_{\mathtt{CA}} + N_{\mathtt{CG}} + N_{\mathtt{TA}} + N_{\mathtt{TG}}}{\pi(\mathtt{A}) + \pi(\mathtt{G})}, \tag{2.14}$$

where $N_{\cdot\cdot} = \sum_{w_1} N_{w_1 \cdot} = N_{\mathtt{M_B}} + N_{\mathtt{M_A}}$ is the total number of matches in the before and after gene states. In the estimation step we replace the count statistic in (2.13) and (2.14) by their conditional mean values given the observed sequences. 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 $N_{\mathtt{M_B}} + N_{\mathtt{M_A}}$ and the conditional mean values of the right hand sides of (2.13) and (2.14). In the Appendix we construct moment equations for parameter estimation in the original HKY-model.

In case of uniform frequencies $\pi(\mathtt{A}) = \pi(\mathtt{G}) = \pi(\mathtt{C}) = \pi(\mathtt{T}) = 1/4$ we get from (2.8), (2.13) and (2.14)

$$s_{\mathtt{B}} = \frac{1}{4} + \frac{\kappa_{\mathtt{B}}}{2}, \quad N_{\cdot\cdot} \frac{\tau_{\mathtt{B}}}{s_{\mathtt{B}}} = 4 N_{\mathtt{ts}}, \quad N_{\cdot\cdot} \kappa_{\mathtt{B}} \frac{\tau_{\mathtt{B}}}{s_{\mathtt{B}}} = 2 N_{\mathtt{tv}},$$

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

$$\tau_{\mathtt{B}} = \frac{N_{\mathtt{ts}} + N_{\mathtt{tv}}}{N_{\cdot\cdot}}, \quad \kappa_{\mathtt{B}} = \frac{N_{\mathtt{tv}}}{2 N_{\mathtt{ts}}}. \tag{2.15}$$

Thus if the nucleotides are uniformly distributed the time span $\tau_{\mathtt{B}}$ is estimated as the fraction of nucleotides undergoing changes and the transversion-transition ratio $\kappa_{\mathtt{B}}$ 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 (2.15) as given by (2.8), (2.13) and (2.14).

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 $\tau_{\mathtt{C}}, \kappa_{\mathtt{C}}, \omega_{\mathtt{C}}, \rho_{\mathtt{C}}, \theta_{\mathtt{C}}$. Now let $N_{w_1 w_2}$ denote the number of codon substitutions of w_1 by w_2 in the match codon state.

From (2.10) we get, with $w_1 \neq w_2$,

$$N_{w_1w_2} = \begin{cases} \tau_{\mathtt{C}}\pi(w_2)N_{w_1\cdot}/s_{\mathtt{C}} & \text{for synonymous transition} \\ \kappa_{\mathtt{C}}\tau_{\mathtt{C}}\pi(w_2)N_{w_1\cdot}/s_{\mathtt{C}} & \text{for synonymous transversion} \\ \omega_{\mathtt{C}}\tau_{\mathtt{C}}\pi(w_2)N_{w_1\cdot}/s_{\mathtt{C}} & \text{for nonsynonymous transition} \\ \rho_{\mathtt{C}}\tau_{\mathtt{C}}\pi(w_2)N_{w_1\cdot}/s_{\mathtt{C}} & \text{for nonsynonymous transversion} \\ \theta_{\mathtt{C}}\tau_{\mathtt{C}}^2\pi(w_2)N_{w_1\cdot}/s_{\mathtt{C}} & \text{otherwise.} \end{cases}$$

Adding e.g. all synonymous transversion equalities we obtain

$$N_{\cdot\cdot}\frac{\tau_{\mathtt{C}}}{s_{\mathtt{C}}} = \sum_{w_1} \frac{\sum_{w_2} N_{w_1w_2} 1_{\mathtt{s,ts}}(w_1, w_2)}{\sum_{w_2} \pi(w_2) 1_{\mathtt{s,ts}}(w_1, w_2)} = N_{\mathtt{s,ts}}, \tag{2.16}$$

where $1_{\mathtt{s,ts}}(w_1, w_2)$ 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 $N_{\cdot\cdot} = N_{\mathtt{M_C}}$ and of the right hand sides of the five equations similar to (2.16). 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 Figure 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 $l(\mathtt{M_B}) = (1, 1)$, $l(\mathtt{D_B}) = (1, 0)$ and $l(\mathtt{M_C}) = (3, 3)$. Then the recursion is

$$P(K|x) = \sum_y p(x, y)p_e(K, l(y)|y)P(K + l(y)|y), \tag{2.17}$$

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 $P(1, 1|\mathtt{Begin})$. Note that the sum in (2.17) 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

$$P(L + 1|x) = p(x, \mathtt{End}), \tag{2.18}$$

where `End` is the state shown in Figure 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 Figure 1. Also let $S[x_i] = (S_1[x_i], S_2[x_i])$ be the nucleotides emitted by x_i in the sequence $x_1, \ldots, x_n$. 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

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

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

$$\frac{p(x, y)p_e(K, l(y)|y)P(K + l(y)|y)}{P(K|x)}.$$

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

$$\begin{aligned} EN_A(K|x) = \; & \textstyle\sum_y \; \Big(1_A(y, S[y]) + EN_A(K + l(y))\Big) \times \\ & \qquad \frac{p(x, y)p_e(K, l(y)|y)P(K + l(y)|y)}{P(K|x)}. \end{aligned} \tag{2.19}$$

The start of the recursion is given by

$$EN_A(L + 1|x) = 0.$$

2.4 Application to *A.tumefaciens* and *M.loti*

We applied the pair prokaryotic HMM to analyse two homologous sequences from *Agrobacterium tumefaciens* and *Mesorhizobium loti*. Genbank accession numbers are AE009042 and AP003011. The sequences code for the protein AGR_C_1356p, which is a exodeoxyribonuclease small subunit. The modified EM-algorithm described in Section 2.3 based on moment equations was used for parameter estimation. 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

	l	$\tau_{\mathtt{B}}$	$\kappa_{\mathtt{B}}$	$\gamma_{\mathtt{B}}$	$\mu_{\mathtt{B}}$
Start	-769.964	0.400	0.500	0.995	0.250
Iteration 1	-719.513	0.438	0.750	0.994	0.100
Iteration 2	-713.171	0.454	0.837	0.994	0.052
Iteration 3	-712.383	0.458	0.855	0.994	0.040
Iteration 5	-712.282	0.458	0.859	0.994	0.036
Iteration 10	-712.279	0.458	0.859	0.994	0.036
Iteration 10	-715.450	0.458	0.859	0.994	0.037

	$\tau_{\mathtt{C}}$	$\kappa_{\mathtt{C}}$	$\omega_{\mathtt{C}}$	$\rho_{\mathtt{C}}$	$\theta_{\mathtt{C}}$	$\gamma_{\mathtt{C}}$	$\mu_{\mathtt{C}}$
Start	0.300	0.500	0.400	0.300	0.300	0.995	0.250
Iteration 1	0.542	0.381	0.126	0.228	0.104	0.988	0.015
Iteration 2	0.544	0.379	0.124	0.227	0.104	0.988	0.011
Iteration 3	0.544	0.379	0.124	0.227	0.104	0.988	0.011
Iteration 5	0.544	0.379	0.124	0.227	0.104	0.988	0.011
Iteration 10	0.544	0.379	0.124	0.227	0.104	0.988	0.011
Iteration 10	0.492	0.409	0.187	0.076	0.115	0.988	0.012

Table 5: EM-algorithm for the pair prokaryotic HMM. The parameter values shown at several iterations are from the full Goldman and Yang model. The parameter values shown after 10 iterations only are from the constrained Goldman and Yang model.

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 Figure 2 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 $\rho_{\mathtt{C}} = \kappa_{\mathtt{C}}\omega_{\mathtt{C}}$ fits the data. The parameters of the constrained Goldman and Yang model are estimated as follows. Recall that the five substitution parameters $\tau_{\mathtt{C}}, \kappa_{\mathtt{C}}, \omega_{\mathtt{C}}, \rho_{\mathtt{C}}, \theta_{\mathtt{C}}$, in the coding part of the sequences are estimated from five equations of the type (2.16). 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 hand sides of these equations. Letting $EN_{\mathtt{s,ts}}, EN_{\mathtt{s,tv}}, EN_{\mathtt{ns,ts}}, EN_{\mathtt{ns,tv}}, EN_{\mathtt{other}}$ denote the counts on the right hand sides (with obvious notation) the estimates in each iteration minimize the sum of squares

$$(\frac{EN_{\mathtt{M_c}}\tau_{\mathtt{C}}}{s_{\mathtt{C}}} - EN_{\mathtt{s,ts}})^2 + (\frac{EN_{\mathtt{M_c}}\kappa_{\mathtt{C}}\tau_{\mathtt{C}}}{s_{\mathtt{C}}} - EN_{\mathtt{s,tv}})^2 + (\frac{EN_{\mathtt{M_c}}\omega_{\mathtt{C}}\tau_{\mathtt{C}}}{s_{\mathtt{C}}} - EN_{\mathtt{ns,ts}})^2 + \\ (\frac{EN_{\mathtt{M_c}}\kappa_{\mathtt{C}}\omega_{\mathtt{C}}\tau_{\mathtt{C}}}{s_{\mathtt{C}}} - EN_{\mathtt{ns,tv}})^2 + (\frac{EN_{\mathtt{M_c}}\theta_{\mathtt{C}}\tau_{\mathtt{C}}^2}{s_{\mathtt{C}}} - EN_{\mathtt{other}})^2.$$

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

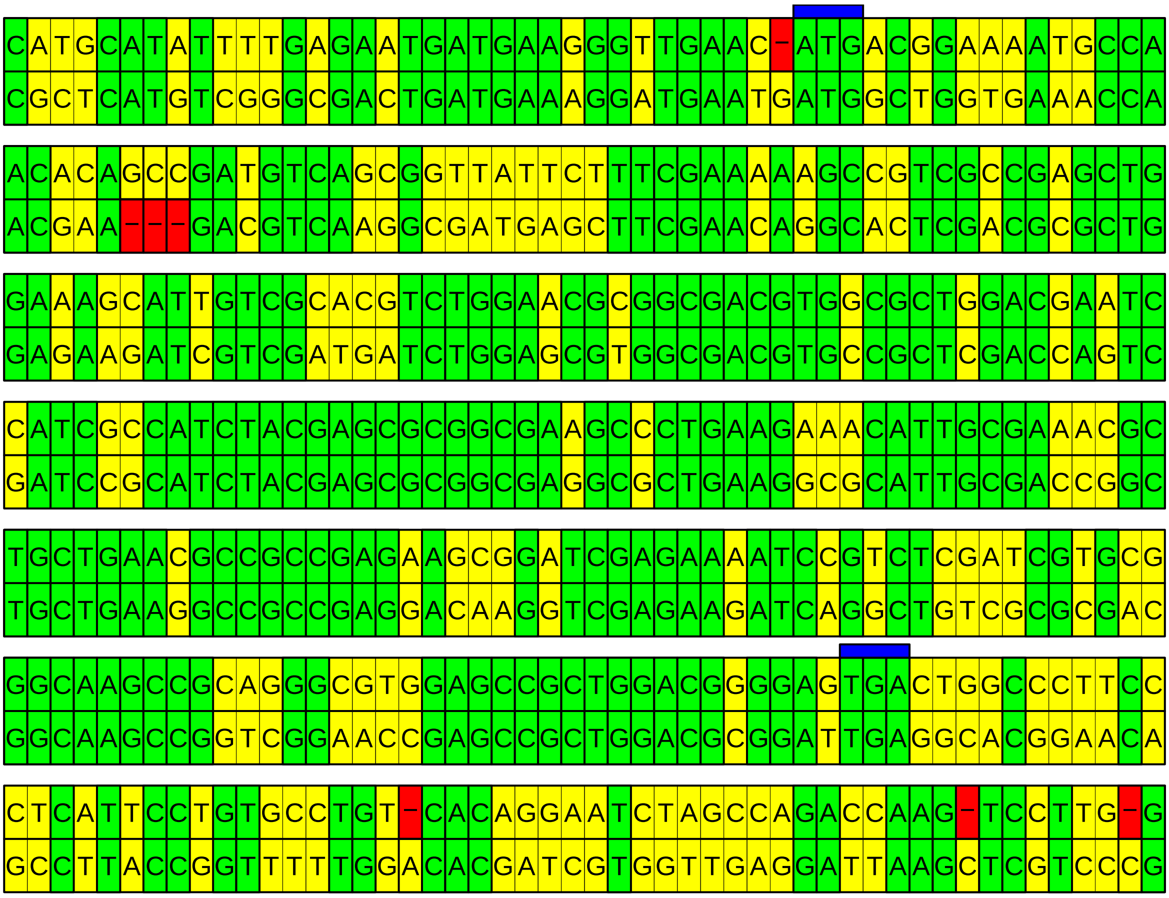
CATGCATATTTTGAGAATGATGAAGGGTTGAAC−ATGACGGAAAATGCCA
CGCTCATGTCGGGCGACTGATGAAAGGATGAATGATGGCTGGTGAAACCA
ACACAGCCGATGTCAGCGGTTATTCTTTCGAAAAAGCCGTCGCCGAGCTG
ACGAA−−−GACGTCAAGGCGATGAGCTTCGAACAGGCACTCGACGCGCTG
GAAAGCATTGTCGCACGTCTGGAACGCGGCGACGTGGCGCTGGACGAATC
GAGAAGATCGTCGATGATCTGGAGCGTGGCGACGTGCCGCTCGACCAGTC
CATCGCCATCTACGAGCGCGGCGAAGCCCTGAAGAAACATTGCGAAACGC
GATCCGCATCTACGAGCGCGGCGAGGCGCTGAAGGCGCATTGCGACCGGC
TGCTGAACGCCGCCGAGAAGCGGATCGAGAAAATCCGTCTCGATCGTGCG
TGCTGAAGGCCGCCGAGGACAAGGTCGAGAAGATCAGGCTGTCGCGCGAC
GGCAAGCCGCAGGGCGTGGAGCCGCTGGACGGGGAGTGACTGGCCCTTCC
GGCAAGCCGGTCGGAACCGAGCCGCTGGACGCGGATTGAGGCACGGAACA
CTCATTCCTGTGCCTGT−CACAGGAATCTAGCCAGACCAAG−TCCTTG−G
GCCTTACCGGTTTTTGGACACGATCGTGGTTGAGGATTAAGCTCGTCCCG

Figure 2: Part of the pairwise alignment of *A.tumefaciens* and *M.loti.* Green colour corresponds to conserved positions, yellow to nonconserved and gaps are shown in red. The two blue bars on top of the alignment indicate the start and stop of the gene.

Carrying out a goodness of fit test of the pair prokaryotic HMM with the constrained Goldman and Yang model (2.10) with $\rho_{\mathtt{C}} = \kappa_{\mathtt{C}}\omega_{\mathtt{C}}$ 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 1 degree of freedom. Using the $\chi^2(1)$ 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

If the two homologous DNA sequences come from eukaryotic organisms we have to introduce intronic parts to the Markov chain depicted in Figure 1. An intronic part can start in three possible phases $0, 1, 2$, 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 Figure 3.

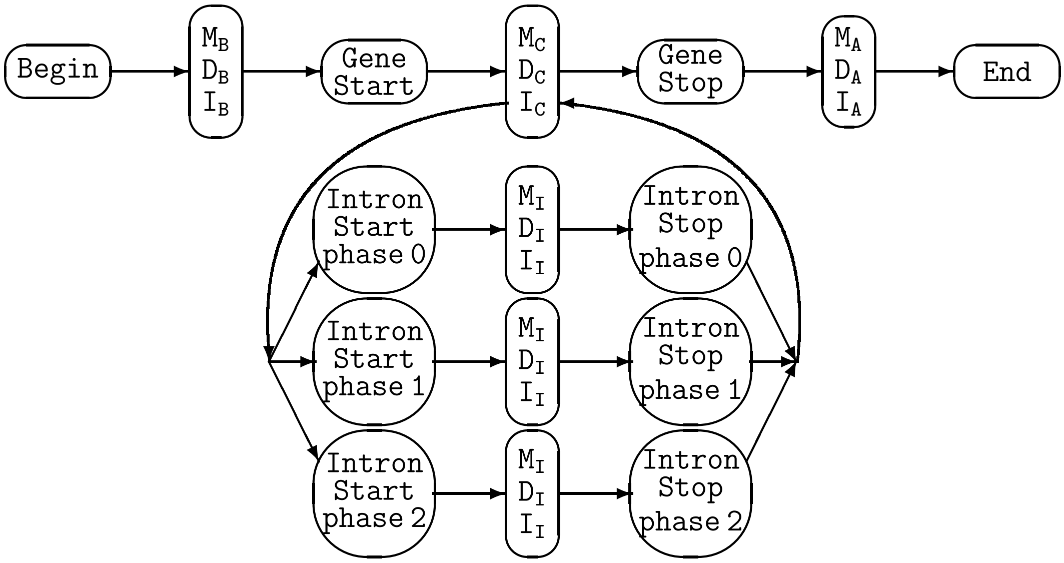

Figure 3: States and transitions of the pair HMM for eukaryotic gene structure prediction.

In Figure 3 the symbols $\mathtt{M_I}, \mathtt{D_I}, \mathtt{I_I}$ 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. Similar 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 at $m/(m+1)$.

Meyer and Durbin (2002) extend the model in Figure 3 by allowing introns within untranslated regions of genes. They also allow introns which are only present in one

IntronStart			IntronStop		
Phase 0	Phase 1	Phase 2	Phase 0	Phase 1	Phase 2
$\begin{pmatrix}\texttt{GT}\\ \texttt{GT}\end{pmatrix}$	$\begin{pmatrix}\texttt{\#GT}\\ \texttt{\#GT}\end{pmatrix}$	$\begin{pmatrix}\texttt{\#\#GT}\\ \texttt{\#\#GT}\end{pmatrix}$	$\begin{pmatrix}\texttt{AG}\\ \texttt{AG}\end{pmatrix}$	$\begin{pmatrix}\texttt{AG\#\#}\\ \texttt{AG\#\#}\end{pmatrix}$	$\begin{pmatrix}\texttt{AG\#}\\ \texttt{AG\#}\end{pmatrix}$

Table 6: Letters emitted from the intron start and stop states, taking into account that introns can come in three different phases depending on the reading frame of the previous exon.

of the genes. Further extensions of the model include e.g. sequencing errors and signals such as the TATA box in the promotor region of the gene and the Poly-A signal at the end of transcription, see Zhang (1998).

The modified EM-algorithm has been applied to several homologous sequences from eukaryotic organisms. In all cases the EM-algorithm converges to a maximum in a few iterations.

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. Again we use a hidden Markov model along the sequences to describe the evolutionary relationship of the sequences. If the model is time-reversible the three sequences are related in a 3-star tree with the observed sequences at the leaves and an unobserved common ancestral sequence in the interior node. Thus for a 3-star tree the hidden states are alignment columns with 4 entries, the first corresponding to the interior node. Symbolically we write a hidden state x as $x = (x_0|x_1, x_2, x_3)$.

The number of emitted nucleotides in the different parts of the Markov chain follow the same rules as in Table 1. In the before and after gene states each entry $x_j, j = 0, 1, 2, 3$, emits a single nucleotide or a gap, $x_j \in \{\#, -\}$, and in the inside gene state each entry emits a sense codon or a gap triplet, $x_j \in \{\#\#\#, - - -\}$. Therefore each of these three parts of the hidden Markov chain have 15 hidden states since the state with gaps in all entries is excluded. We denote the set of 15 states Ω.

In Figure 4 the HMM for triplewise prokaryotic gene structure prediction is depicted. The only difference compared to the HMM for pairwise prokaryotic gene structure prediction as shown in Figure 1 is that alignment columns with 4 entries are emitted instead of alignment columns with 2 entries.

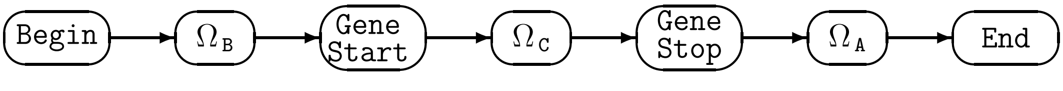

Figure 4: States and transitions of the triple HMM for prokaryotic gene structure prediction. Here the set Ω consists of 15 hidden states generalizing the 3 (match, delete and insert) hidden states of the pair HMM.

4.1 Transition probabilities

We now describe how to extend the TKF-model of Section 2.3 to a 3-star tree. Denote the evolutionary times along the branches of the 3-star tree τ_1, τ_2 and τ_3. Further suppose the parameters of the birth and death process μ and λ are the same along the branches. To state the transition probabilities we define $\beta_j, b_j(\cdot,\cdot), s_j(\cdot)$, as in (2.1)-(2.4) with τ replaced by $\tau_j, j = 1, 2, 3$.

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 γ 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.

	$y_0 = \#$	$y_0 = -$	$y = \texttt{End}$
$x_0 = \#$	$\left\{\prod_{j=1}^{3} b_j(x_j,-)\right\}\gamma\prod_{j=1}^{3} s_j(y_j)$	$\prod_{j=1}^{3} b_j(x_j,y_j)$	$\left\{\prod_{j=1}^{3} b_j(x_j,-)\right\}(1-\gamma)$
$x_0 = -$	$\left\{\prod_{\{j\geq 1:x_j=\#\}} b_j(\#,-)\right\}\gamma\prod_{j=1}^{3} s_j(y_j)$	$\prod_{\{j\geq 1:x_j=\#\}} b_j(\#,y_j)$	$\left\{\prod_{\{j\geq 1:x_j=\#\}} b_j(\#,-)\right\}(1-\gamma)$

Table 7: Transition probabilities in the 3-star TKF-model. The terms $b_j(\cdot,\cdot)$ and $s(\cdot)$ are defined in (2.1)-(2.4) with τ replaced by $\tau_j, j = 1, 2, 3$. Transitions from a state with $x_0 = -$ to a state with $y_0 = -$ is only possible if $y_j = -$ when $x_j = -$.

As in the case of two sequences 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 in Table 7 with $x = (\#|\#, \#, \#)$.

4.2 Emission probabilities

Recall that a hidden state x is an alignment column with 4 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 $x_j = -$. Following (2.6) the emission probabilities in the before and after gene states are given by

$$p_e^0(w|x) = \begin{cases} \pi(w_0) \displaystyle\prod_{\{j\geq 1:x_j=\#\}} f_j(w_j|w_0) & \text{if } x_0 = \# \\ \displaystyle\prod_{\{j\geq 1:x_j=\#\}} \pi(w_j) & \text{if } x_0 = -. \end{cases} \tag{4.1}$$

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

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

$$p_e((w_1, w_2, w_3)|x) = \begin{cases} \sum_{w_0} \pi(w_0) \prod_{\{j\geq 1: x_j = \#\}} f(w_j|w_0) & \text{if } x_0 = \# \\ \prod_{\{j\geq 1: x_j = \#\}} \pi(w_j) & \text{if } x_0 = -, \end{cases} \tag{4.2}$$

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

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

since π 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

$$p_e^0(w_0|(w_1, w_2, w_3), x) = \frac{p_e^0(w|x)}{p_e((w_1, w_2, w_3)|x)}, \tag{4.3}$$

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

The emission probabilities in the inside gene states are defined as in (4.1) with π replaced by $\pi_{\texttt{C}}$ and with $f_j(\cdot, \cdot)$ determined by (2.10). Here the evolutionary distance $\tau_{\texttt{C}}$ and the parameters $\kappa_{\texttt{C}}, \omega_{\texttt{C}}, \rho_{\texttt{C}}, \theta_{\texttt{C}}$ 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

$$p_e^0(w|x) = \pi_{\texttt{S}}(w_0) \prod_{j=1}^{3} f_j(w_j|w_0), \quad w_j \in \{\texttt{TAA}, \texttt{TAG}, \texttt{TGA}\}, \quad j = 0, 1, 2, 3.$$

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 $\gamma_{\texttt{B}}, \mu_{\texttt{B}}, \gamma_{\texttt{C}}, \mu_{\texttt{C}}$ be common parameters on all lineages. For the remaining parameters we consider the full model with $\tau_{\texttt{B}}, \kappa_{\texttt{B}}, \tau_{\texttt{C}}, \kappa_{\texttt{C}}, \omega_{\texttt{C}}, \rho_{\texttt{C}}, \theta_{\texttt{C}}$ 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 $N_{\texttt{A}_{\texttt{C}}}$ be the number of states having the symbol ### in the ancestral sequence and $N_{\texttt{FM}_{\texttt{C}}}$ be the number of full matches (###|###, ###, ###). From (2.5) we can write the moment equations

$$N_{\texttt{A}_{\texttt{C}}} = \gamma_{\texttt{C}}/(1 - \gamma_{\texttt{C}}), \quad N_{\texttt{FM}_{\texttt{C}}} = \exp\Big(-\mu_{\texttt{C}}(\tau_{\texttt{C},1} + \tau_{\texttt{C},2} + \tau_{\texttt{C},3})\Big) N_{\texttt{A}_{\texttt{C}}}. \tag{4.4}$$

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 $\mathtt{Q_B} = (\#|-,-,-), \mathtt{Q_A} = (\#|-,-,-)$ and $\mathtt{Q_C} = (\#\#\#|---,---,---)$ the start of the recursion (2.18) and the recursion (2.17) become more complicated. With $(L_1, L_2, L_3) = L$ the start of the recursion now becomes

$$P(L+1|\mathtt{Q_A}) = \frac{p(\mathtt{Q_A}, \mathtt{End})}{1 - p(\mathtt{Q_A}, \mathtt{Q_A})},$$

and for $x \neq \mathtt{Q_A}$,

$$P(L+1|x) = p(x, \mathtt{End}) + p(x, \mathtt{Q_A}) P(L+1|\mathtt{Q_A}),$$

where End is the state shown in Figure 4. The recursion is given by first finding the marginal probability of the sequences $S[K : L+1]$ given the initial state Q is one of the silent states $\mathtt{Q_B}, \mathtt{Q_A}, \mathtt{Q_C}$

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

and second finding the marginal probability for the non-silent states as in (2.17).

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 (2.19) by

$$\sum_{w_0} 1_A(y, w_0, S[y]) p_e^0(w_0|S[y], y).$$

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

We applied the 3-star prokaryotic HMM to analyse 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. 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 Figure 5 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 $\kappa_\mathtt{B}$ and $\kappa_\mathtt{C}$ to be the same in all branches, and perhaps even the same in the intergenic and coding parts. The synonymous-nonsynonymous ratio $\omega_\mathtt{C}$ is of interest on its own since a value of $\omega_\mathtt{C}$ 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

	$\tau_{\mathtt{B}}$	$\kappa_{\mathtt{B}}$	$\tau_{\mathtt{C}}$	$\kappa_{\mathtt{C}}$	$\omega_{\mathtt{C}}$	$\rho_{\mathtt{C}}$	$\theta_{\mathtt{C}}$
A. tumefaciens	0.188	0.629	0.315	0.257	0.010	0.076	0.063
M.loti	0.331	0.855	0.367	0.765	0.397	0.632	0.386
S.meliloti	0.271	0.807	0.371	0.266	0.002	0.001	0.026

$$\gamma_{\mathtt{B}} = 0.994,\ \mu_{\mathtt{B}} = 0.064,\ \gamma_{\mathtt{C}} = 0.988,\ \mu_{\mathtt{C}} = 0.002,\ l = -2128.5.$$

Table 8: EM-algorithm for the triple prokaryotic HMM.

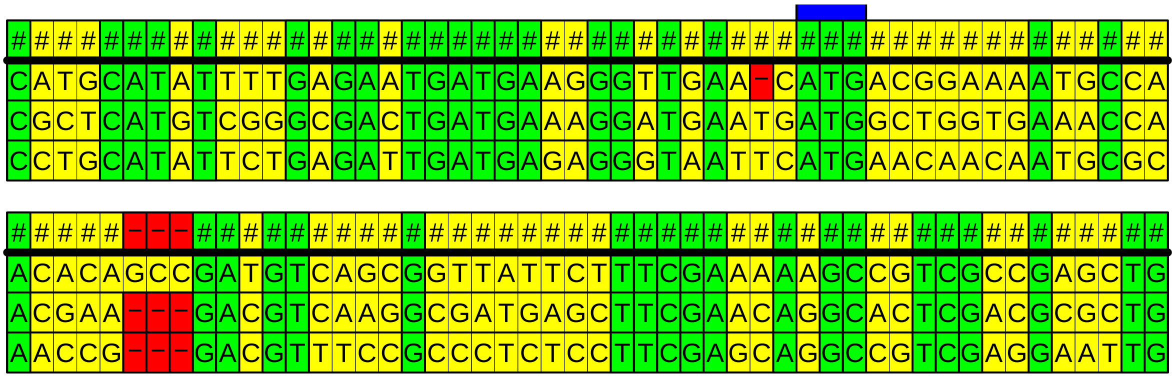
##
CATGCATATTTTGAGAATGATGAAGGGTTGAA–CATGACGGAAAATGCCA
CGCTCATGTCGGGCGACTGATGAAAGGATGAATGATGGCTGGTGAAACCA
CCTGCATATTCTGAGATTGATGAGAGGGTAATTCATGAACAACAATGCGC
#####–––##
ACACAGCCGATGTCAGCGGTTATTCTTTCGAAAAAGCCGTCGCCGAGCTG
ACGAA–––GACGTCAAGGCGATGAGCTTCGAACAGGCACTCGACGCGCTG
AACCG–––GACGTTTCCGCCCTCTCCTTCGAGCAGGCCGTCGAGGAATTG

Figure 5: Part of the multiple alignment of *A.tumefaciens*, *M.loti* and *S.meliloti*. The top row specifies the hidden state of the ancestral sequence.

the evolutionary distances to scale linearly in the intergenic and coding regions such that

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

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

We fitted the constrained model with $\kappa_{\mathtt{B}}$ being the same in all branches. The fitted value of $\kappa_{\mathtt{B}}$ is 0.788, and the remaining parameter values only changed 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 2 degrees of freedom. Using the $\chi^2(2)$ approximation of the test statistic the p-value is 82%, and thus support the expectation that the constrained model is sufficiently flexible compared to the full model.

5 Discussion

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 $O(NL_1L_2)$ and $O(TL_1L_2L_3)$. For two sequences Meyer and Durbin (2002) have developed the stepping stone algorithm, where subsequences of strong similarity are used as fixed points for the alignment. A similar algorithm can be formulated for multiple sequences and is needed if alignment and gene structure prediction are carried out simultaneously. Another approach is to search the alignment space using simulation procedures as discussed by Holmes and Bruno (2001) and Jensen and Hein (2002).

Acknowledgement

We thank Jacob S. Pedersen for sharing his knowledge on gene finding and Leif Schauser for providing the data.

References

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78-94.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *J. Mol. Evol.* **22**, 725-735.

Hasegawa, M., Kishino, H. and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160-174.

Hein, J., Jensen, J.L. and Pedersen, C.N.S. (2003). Recursions for statistical multiple alignment. Research Report no. 425. Department of Theoretical Statistics, University of Aarhus. To appear in *Proc. Natl. Acad. Sci. USA*.

Holmes, I. and Bruno, W.J. (2001). Evolutionary HMMs: A Bayesian approach to multiple alignment. *Bioinformatics*, **17**, 803-820.

Jensen, J.L. and Hein, J. (2002). Gibbs sampler for statistical multiple alignment. Research Report no. 429. Department of Theoretical Statistics, University of Aarhus.

Krogh, A. (1997). Two methods for improving performance of a HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 179-186.

Meyer, I.M. and Durbin, R. (2002). Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics*, **18**, 1309-1318.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929-936.

Pachter, L., Alexandersson, M. and Cawley, S. (2002). Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J. Comp. Biol.* **9**, 389-399.

Pedersen, J.S. and Hein, J. (2003). Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics*, **19**, 219-227.

Thorne, J.L., Kishino, H. and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33**, 114-124.

Zhang, M.Q. (1998). Statistical features of human exons and their flanking regions. *Human Molecular Genetics*, **7**, 919-932.

Appendix

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

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

$$Q(w_1, w_2) = \begin{cases} \tau_{\mathtt{B}}\pi(w_2)/s_{\mathtt{B}} & \text{for transition} \\ \kappa_{\mathtt{B}}\tau_{\mathtt{B}}\pi(w_2)/s_{\mathtt{B}} & \text{for transversion,} \end{cases}$$

for $w_1 \neq w_2$, with corresponding substitution probabilities given by the matrix $\exp(Q\tau_{\mathtt{B}})$. Let N_w denote the number of times w occur in sequence S_1 in the before or after match states. Further let $N_{w_1w_2}$ 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 $\tau_{\mathtt{B}}$ and $\kappa_{\mathtt{B}}$ from the two moment equations

$$\begin{aligned} & N_{\mathtt{AG}} + N_{\mathtt{GA}} + N_{\mathtt{CT}} + N_{\mathtt{TC}} = \\ & \quad N_{\mathtt{A}}p_{\mathtt{AG}}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}}) + N_{\mathtt{G}}p_{\mathtt{GA}}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}}) + N_{\mathtt{C}}p_{\mathtt{CT}}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}}) + N_{\mathtt{T}}p_{\mathtt{TC}}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}}) \end{aligned} \tag{5.1}$$

$$\begin{aligned} & N_{\mathtt{AC}} + N_{\mathtt{AT}} + N_{\mathtt{GC}} + N_{\mathtt{GT}} + N_{\mathtt{CA}} + N_{\mathtt{CG}} + N_{\mathtt{TA}} + N_{\mathtt{TG}} = \\ & \quad N_{\mathtt{A}}(p_{\mathtt{AC}}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}}) + p_{\mathtt{AT}}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}})) + N_{\mathtt{G}}(p_{\mathtt{GC}}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}}) + p_{\mathtt{GT}}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}})) + \\ & \quad N_{\mathtt{C}}(p_{\mathtt{CA}}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}}) + p_{\mathtt{CG}}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}})) + N_{\mathtt{T}}(p_{\mathtt{TA}}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}}) + p_{\mathtt{TG}}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}})), \end{aligned} \tag{5.2}$$

where $p_{w_1w_2}(\tau_{\mathtt{B}}, \kappa_{\mathtt{B}})$ is the $(w_1, w_2)'$th entry in $\exp(Q\tau_{\mathtt{B}})$. The equations would have to be solved numerically and require six counts, namely $N_{\mathtt{A}}, N_{\mathtt{G}}, N_{\mathtt{C}}, N_{\mathtt{T}}$ and the left hand sides of (5.1) and (5.2).

In the Goldman and Yang model (2.9) with $\kappa_{\mathtt{C}}\omega_{\mathtt{C}}$ replaced by the free parameter $\rho_{\mathtt{C}}$ we get with a similar notation

$$\sum_{w_1,w_2} N_{w_1w_2} 1_{\mathtt{s,ts}}(w_1, w_2) = \sum_{w_1,w_2} N_{w_1} p_{w_1w_2}(\tau_{\mathtt{C}}, \kappa_{\mathtt{C}}, \omega_{\mathtt{C}}, \rho_{\mathtt{C}}) 1_{\mathtt{s,ts}}(w_1, w_2),$$

where $1_{\mathtt{s,ts}}(w_1, w_2)$ is 1 if the change from w_1 to w_2 is a synonymous transition and 0 otherwise. Similarly three other equations with $1_{\mathtt{s,ts}}$ replaced by $1_{\mathtt{s,tv}}$ (synonymous transversions), $1_{\mathtt{ns,ts}}$ (nonsynonymous transitions), and $1_{\mathtt{ns,tv}}$ (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 hand sides.

Asger Hobolth, Bioinformatics Research Centre, Department of Computer Science,
University of Aarhus, Ny Munkegade, DK-9000 Aarhus C, Denmark.
E-mail: asger@birc.dk.