



## Gene finding with a hidden Markov model of genome structure and evolution

Jakob Skou Pedersen<sup>1,\*</sup> and Jotun Hein<sup>2</sup>

<sup>1</sup>Bioinformatics Research Center, Department of Genetics and Ecology, The Institute of Biological Sciences, University of Aarhus, Building 550, Ny Munkegade, 8000 Aarhus C, Denmark and <sup>2</sup>Department of Statistics, Oxford University, The Peter Medawar Building for Pathogen Research, South Parks Road, Oxford OX1 3SY, UK

Received on March 28, 2002; revised on June 3, 2002; accepted on July 17, 2002

### ABSTRACT

**Motivation:** A growing number of genomes are sequenced. The differences in evolutionary pattern between functional regions can thus be observed genome-wide in a whole set of organisms. The diverse evolutionary pattern of different functional regions can be exploited in the process of genomic annotation. The modelling of evolution by the existing comparative gene finders leaves room for improvement.

**Results:** A probabilistic model of both genome structure and evolution is designed. This type of model is called an Evolutionary Hidden Markov Model (EHMM), being composed of an HMM and a set of region-specific evolutionary models based on a phylogenetic tree. All parameters can be estimated by maximum likelihood, including the phylogenetic tree. It can handle any number of aligned genomes, using their phylogenetic tree to model the evolutionary correlations. The time complexity of all algorithms used for handling the model are linear in alignment length and genome number. The model is applied to the problem of gene finding. The benefit of modelling sequence evolution is demonstrated both in a range of simulations and on a set of orthologous human/mouse gene pairs.

**Availability:** Free availability over the Internet on www server: <http://www.birc.dk/Software/evogene>

**Contact:** [jsp@daimi.au.dk](mailto:jsp@daimi.au.dk)

### INTRODUCTION

The genomic sequencing projects are moving into a new era. In recent years complete genomes from diverse eukaryotic organisms have been sequenced. The sequencing of closely related genomes both between and within organisms has now begun. This allows interesting comparative studies, and integration of evolutionary information in structural genomic analysis.

Evolutionary information in the form of homology

searches has long been used by the gene-finding community (Staden and McLachlan, 1982). The success of these searches is limited both by the presence of homologs in the query database, and by the database quality with regard to correct annotations. Alignment of homologous genomes will make it possible to analyze the evolutionary process genome-wide. The diverse selection pressure between different functional regions results in correspondingly diverse evolutionary patterns. Information on the evolutionary pattern can therefore be used to characterize not just coding regions—but all different functional regions. The predictive power of gene finders can thereby be increased, without excluding the use of traditional homology searches. Some recent improvements in this direction have been made. They can be grouped into methods which align and annotate in one combined procedure, and methods which rely on pre-produced alignments. Both types rely on the conservation of gene structure even for considerably divergent sequences (Batzoglou *et al.*, 2000).

Pachter *et al.* (2001) describe a framework called a Generalized Pair Hidden Markov Model (GPHMM), which combines the general HMMs used in gene finding (Kulp *et al.*, 1996; Burge and Karlin, 1997) with the pair HMMs used for alignment (Durbin *et al.*, 1998). The Conserved Exon Method (Bafna and Huson, 2000) finds potential exons pairs between two genomes by scoring their alignment. They mention, but do not state, recursions for aligning and annotating genes in genomic sequences. Such recursions have only been explicitly stated in unpublished work (Blayo *et al.*, 1999; Scharling, 2001).

The two-step method GLASS and ROSETTA (Batzoglou *et al.*, 2000) aligns co-linear homologous genomic sequence and scores exons according to a range of criteria, including amino acid similarity and splice site quality. TWINSKAN (Korf *et al.*, 2001) extends GENSCAN (Burge and Karlin, 1997) by including a component for genome comparison. It uses a blast version to search

\*To whom correspondence should be addressed.

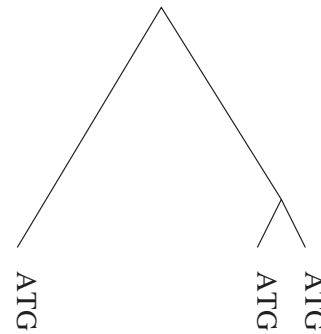
the query genome against an organism specific-genomic database. From this an aligned sequence of match, mismatch and misalignment symbols is made against the query genome. The pattern of this sequence is modelled for different genomic regions. Nekrutenko *et al.* (2001) have made an exon classifier based on pairwise alignments of putative exons. The classifier tests if the putative exon pair exhibits a significant deviation from one in its ratio of non-synonymous to synonymous substitutions (the  $dN/dS$ -ratio, see below).

The above-mentioned methods do not exploit all the information available in the evolutionary pattern, neither are they easily extended to multiple genomes. Here the use of an HMM for modelling gene structure and continuous Markov chains for modelling the evolutionary pattern of different regions is investigated. HMMs were applied to biological sequence analysis in 1989 (Churchill, 1989), a few years later they were adopted by the gene-finding community (Krogh *et al.*, 1994), where they are now widely used (Kulp *et al.*, 1996; Burge and Karlin, 1997; Krogh, 1997, etc.). Continuous Markov chains are well known as models of molecular evolution in phylogenetic analysis (Liò and Goldman, 1998). This model construction is termed an Evolutionary Hidden Markov Model (EHMM). In the form presented here it relies on a pre-produced alignment.

The benefit of using an EHMM for gene finding is investigated with a minimalistic model of eukaryotic gene structure. The scope of this model is not to compete with existing gene finders on performance, but to illustrate the power of this approach. Each state of the EHMM corresponds to a specific genomic region or signal, containing a parameterized model of the relevant evolutionary process. The evolutionary process is modelled on the phylogenetic tree relating the genomes of the alignment.

The EHMM has a range of properties which makes it well suited for gene finding: (1) when used on a single sequence it has the properties of a traditional, possibly, higher-order HMM; (2) it can handle any number of sequences in the alignment; (3) it can handle variability in the number of sequences along the alignment; (4) state of the art evolutionary models can be incorporated; and (5) evolutionary events between different genomes are not treated independently, but conditioned by their evolutionary relationship through the use of a phylogenetic tree (see Figure 1).

Our work is inspired by Goldman *et al.* (1996), who used an EHMM method for protein secondary structure prediction, i.e. classifying protein positions into either alpha helix, beta sheet or loop structure. Their EHMM contained amino acid level evolutionary models optimized for the three structure types. Knudsen and Hein (1999) used this strategy in the context of stochastic context-free grammars for RNA secondary-structure prediction.



**Fig. 1.** A hypothetical phylogenetic tree of three genomes. Only three homologous nucleotides from each genome are shown. The two sequences to the right are closely related and should not be taken as independent evidence of a start codon.

## THE MODEL

The structure of biological sequences is often modelled by HMMs, and their evolution by continuous Markov chains. An EHMM combines these two model types thereby modelling sequence evolution and structure at the same time. The input consists both of an alignment of sequences and of their relating phylogenetic tree. EHMMs are very flexible, and can extend the HMMs used in many biological applications. The following description will be based on the simple model of eukaryotic genome structure and evolution used for gene finding.

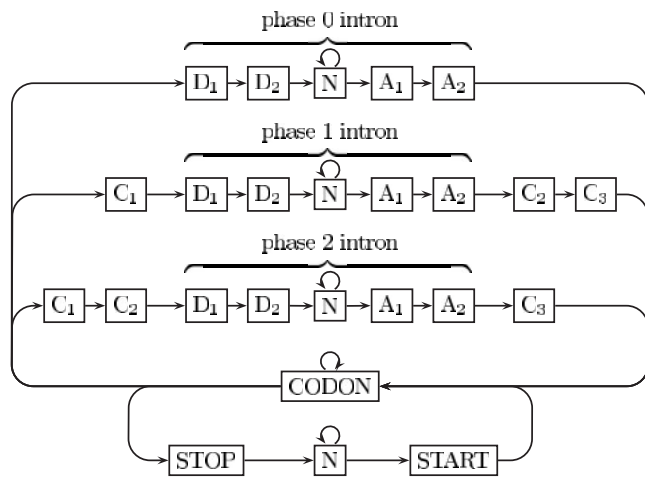
### The hidden Markov model

HMMs can be seen as stochastic regular grammars capable of both generating and parsing sequences from some alphabet (Durbin *et al.*, 1998). HMMs are specified by five components: a set of states, a matrix of transition probabilities ( $A = \{a_{kl}\}$ ), a set of alphabets ( $C$ ), a set of emission distributions ( $e$ ), and an initial state distribution ( $B$ ) (Rabiner, 1989).

When a state  $k$  is visited it emits an observable  $c$  from the alphabet  $C_k$  according to the distribution  $e_k$ . A path of visited states  $\pi = \{\pi_i\}_{i=1}^{i=L}$  of length  $L$  is found by choosing the first state according to  $B$  followed by  $L-1$  state transitions according to  $A$ . The probability of a path  $\pi$  and its corresponding sequence of observables ( $x = \{x_i\}_{i=1}^{i=L}$ ) is:

$$P(x, \pi | A, B, e) = B(\pi_1) e_{\pi_1}(x_1) \prod_{i=2}^L e_{\pi_i}(x_i) a_{\pi_{i-1}\pi_i}. \quad (1)$$

Each state of the eukaryotic genome model corresponds to either a genomic region or is part of a regulatory site. The states and the allowed transitions are shown in



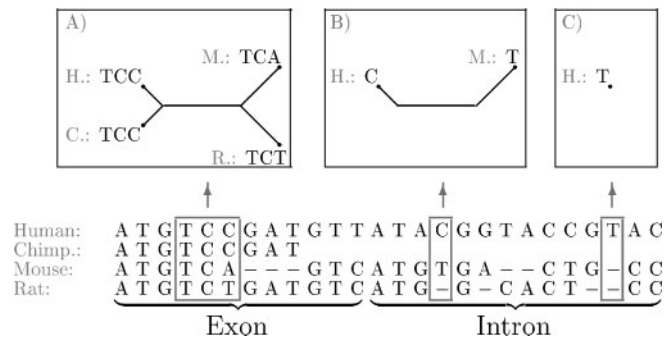
**Fig. 2.** Architecture of eukaryotic EHMM. Each box of the graph is a state of the model, some states are of the same type. Start, stop, and codon denotes start codon, stop codon, and inner codon, respectively. N denotes intergenic/intron state.  $C_1$ ,  $C_2$ , and  $C_3$  denote single-codon positions, in combination they model the codons which are split by introns.  $D_1$ , and  $D_2$  denote first and second position of the splice donor site.  $A_1$ , and  $A_2$  denote first and second position of the splice acceptor site. When an alignment is generated from the model a start state, e.g. the intergenic state, is chosen according to the initial state distribution  $B$ . State transitions, e.g. from inner codon back again to inner codon, are chosen from the matrix of transition probabilities  $A$ . Only transitions with positive probability are shown on the graph. For each state visited an alignment column is generated, e.g. a column of ATGs from the start codon state, according to the state specific emission distribution  $e_k$ . A single exon gene can thus be modelled as follows: start in the intergenic state (the lower N state), make a given number of self transitions, make a transition to the start codon state, make a transition to the inner codon state, make a given number of self transitions, make a transition to the stop codon state, make a transition to the intergenic state. For each state visited the relevant alignment column is emitted. Multi exon genes are modelled by including transitions from the inner codon state to the states in the upper half of the figure. The single-codon-position states maintain the reading frame across introns (Krogh, 1998). This independent treatment of each position in codons split by introns introduces a small risk of producing in frame stop-codons. Only genes on the forward strand are modelled, extending to both strands is simple (cf. Burge and Karlin, 1997).

Figure 2. The model only contains the minimal number of states necessary for modelling eukaryotic genes.

### Alphabets and emission distributions

An EHMM is characterized by every state  $k$  having an alphabet  $C_k$  over alignment columns, and an emission distribution  $e_k$  specified by a state-specific evolutionary model  $E_k$  and a phylogenetic tree  $T$ .

The evolutionary process along the branches of the phy-



**Fig. 3.** The lower part shows an alignment of four hypothetical sequences from human, chimpanzee, mouse and rat. The upper part shows the phylogenetic tree effectively relating the entries of three specific alignment columns. (A) Shows the tree relating a triplet column with no gaps and no missing data. This tree is used in a triplet state in connection with a specific codon-based evolutionary model to calculate the emission probability of the given column. (B) Shows a single-nucleotide column in which the chimpanzee data is missing. The column also includes a gap. The single-nucleotide states will use a nucleotide-based evolutionary model and a reduced phylogenetic tree for calculating the emission probability of the given column. The reduced tree used will miss the branches leading to the missing data and the gap entries. (C) Shows a single-nucleotide column containing 2 gap entries and one with missing data. The resulting tree will therefore consist of a single node. The emission probability of the single symbol connected to the node will be taken from the equilibrium distribution of the relevant evolutionary model.

logenetic tree is modelled by a state-specific continuous Markov chain ( $E_k$ ) on the discrete state space of sequence characters (single nucleotides or triplets, see below).  $E_k$  is specified by an instantaneous rate matrix ( $Q_k$ ), giving the rate of substitution from a given character ( $a$ ) to another ( $b$ ). The entries of the transition probability matrix ( $P_k(t)$ ) specify the substitution probability from  $a$  to  $b$  given a branch length  $t$  ( $P_k(b|a, t)$ ).  $P_k(t)$  can be derived from  $Q_k$  by matrix exponentiation ( $P_k(t) = \exp(tQ_k)$ ) (e.g. Liò and Goldman, 1998).

$T$  consists of a topology and a set of  $2n - 3$  branch lengths, which together specify the evolutionary relationship between the sequences of the alignment. The entries of an alignment column corresponds to the state of the evolutionary process at the leaves of the phylogenetic tree (see Figure 3). The probability of generating an alignment column in a state  $k$  equals the probability of observing a given character pattern on the leaves of  $T$  given the evolutionary process specified by  $E_k$ :

$$e_k(c) = P(c|E_k, T). \quad (2)$$

This would be a simple calculation if the character states of the inner nodes of  $T$  were known. The probability

of observing the specified evolutionary process down through the tree could be found by calculating  $P_k(t)$  for every branch. But, the character states of inner nodes are missing data, the likelihood must thus be calculated as a sum over all possible configurations of inner nodes. The probability of observing a given character in the root is given by the equilibrium distribution due to an assumption of evolutionary equilibrium. This likelihood calculation can be done in linear time due to a dynamic programming algorithm (Felsenstein, 1981).

The states of the EHMM can be divided into two sets according to the size of the alignment columns in their alphabet. Let  $n$  denote the number of sequences in the alignment. The set of single-nucleotide states (the intron/intergenic state, the three single-codon states, and the four splice-site states) have alphabets over  $n * 1$  columns, while the set of triplet states (the start codon state, the inner codon state, and the stop codon state) have alphabets over  $n \times 3$  columns.

The HKY-model is the evolutionary model used for all single nucleotide states (Hasegawa *et al.*, 1985), and the codon model by Goldman and Yang (Goldman and Yang, 1994; Yang, 2000) the evolutionary model for all triplet states. The equilibrium distribution of the evolutionary process is parameterized in both models, this is a necessary property in order to model splice sites, start codons, and stop codons. The transition/transversion ratio is also parameterized in both. The ratio of non-synonymous to synonymous codon substitutions (the dN/dS-ratio) is modelled in the codon model, which is paramount for describing the difference in evolutionary pattern between coding and non-coding regions of the genome. The overall substitution rate is parameterized in both models.

## ALGORITHMS

### Parameter estimation

The parameters of the EHMM ( $A, B, E, T$ ) are estimated by a combination of Baum–Welch (Rabiner, 1989) and Powell (Press *et al.*, 1992) which are both maximum likelihood methods. The parameters of the evolutionary models  $E$  can be divided into the equilibrium frequencies ( $E_{\text{equ}}$ ) and the remaining evolutionary model parameters ( $E_{\text{evo}}$ ).

### Estimation of $A, B$ , and $E_{\text{equ}}$

Baum–Welch, which is an expectation maximization method often used with HMMs, is used to estimate the matrix of transition probabilities  $A$  and  $E_{\text{equ}}$ . The initial state distribution  $B$  could be estimated by Baum–Welch, but is set to 0.00001 for all states except the intergenic.

In traditional HMM-based gene finders, the alphabets  $C$  are over nucleotides and codons, and the emission

distribution  $e$  give the expected frequency of these. The expectation step of Baum–Welch then estimates the expected number of nucleotides or codons emitted from each state, the expected number of state transitions, and the expected number of times a state is used as initial state. In the maximization step the parameters maximizing the probability of these counts are found.

Baum–Welch does not directly apply for  $e$  of EHMMs, as it is parameterized through  $E$ . The parameters of  $E$  can be estimated by traditional, but computationally costly, estimation procedures. The equilibrium frequencies ( $E_{\text{equ}}$ ) constitute the greater part of  $E$ 's parameters. They can be approximated by averaging the state-specific use of observables over all sequences of the alignment. This is done by counting the observables over all sequences of the alignment in the expectation step of Baum–Welch.

### Estimation of $E_{\text{evo}}$ and $T$

Powell, which is a traditional optimization method, is used for estimating  $E_{\text{evo}}$  and the phylogenetic tree  $T$ . It relies on finding a set of conjugate directions in the parameter space, along which the likelihood maximizations are non-interfering.

The likelihood of an alignment ( $x$ ) given a parameterization of the EHMM can be found from Equation (1) by summing over all possible paths:

$$\begin{aligned} P(x|A, B, e) &= \sum_{\text{all } \pi} P(x, \pi|A, B, e) \\ &= \sum_{\text{all } \pi} P(x, \pi|A, B, E, T). \end{aligned} \quad (3)$$

This calculation can be done in time linear to alignment length by the dynamic programming algorithm Forward (Rabiner, 1989). The maximum likelihood estimates of  $E_{\text{evo}}$  and  $T$  can then be found by Powell as:

$$E_{\text{evo}}^{ML} = \operatorname{argmax}_{E_{\text{evo}}} P(x|A, B, E, T), \quad (4)$$

$$T^{ML} = \operatorname{argmax}_T P(x|A, B, E, T). \quad (5)$$

With just a few aligned sequences it is possible to do an exhaustive search of tree topologies. With more sequences it becomes necessary with either a Branch and Bound method or some heuristic approach (e.g. Felsenstein, 1981).

The above procedures can also be used with annotated data, which was the case with the human–mouse data set. The path through the model is then known and Baum–Welch becomes a simple counting of the known state transitions and observable emissions (Krogh, 1997). The likelihood of a given parametrization is then given directly by Equation (1).

## Gene prediction

The most probable path through the model given the data (the maximum *a posteriori* estimate),

$$\begin{aligned}\pi^{MAP} &= \operatorname{argmax}_{\pi} P(\pi|x, A, B, e) \\ &= \operatorname{argmax}_{\pi} P(\pi, x, A, B, e),\end{aligned}\quad (6)$$

is found by Viterbi's algorithm (Viterbi, 1967). The prediction consists of the genes included in  $\pi$ .

## Alignments

Felsenstein's algorithm treats gaps as missing data. Alignment sequences consisting purely of gaps will not affect the likelihood calculations, the calculations are therefore effectively performed on a reduced tree missing the relevant branch (see Figure 3). This property makes it easy to implement situations in which a sequence can only align in part of its length; the unalignable intervals are just filled with gaps.

Gaps misplaced by the alignment procedure can potentially introduce frame-shifts in coding regions. This risk is avoided by letting one sequence be a gap-less reference. For example, if a human genomic sequence has been aligned against a number of sequences from other organisms, all alignment columns containing a gap in the human sequence are removed.

## DATA SETS

The EHMM is fully probabilistic and can therefore easily be used to simulate data. The eukaryotic genome model, excluding its start, stop and intergenic state (see Figure 2), is used for generating alignments. This reduced model produces only inner exons. A range of simulations were performed using different parameter settings. The ability of the model to correctly predict the exons, given the true parameter settings, were tested.

Batzoglou *et al.* (2000) compiled an annotated data set consisting of 117 orthologous genomic sequence pairs from mouse and human, each containing one complete gene. The sequence pair containing the mouse Fabpi gene was discarded due to an incomplete annotation. The remaining sequence pairs were aligned using the GLASS program (Batzoglou *et al.*, 2000). The final data set was symmetric in the sense that alignments referenced both by human and mouse were made for each sequence pair.

The data set was divided into four subsets, which were used to cross-validate the performance. The parameter estimation of each subset was based on the annotation of the reference sequences. First the annotated version of the Baum–Welch procedure was run, estimating  $A$  and  $E_{\text{equ}}$ , secondly  $E_{\text{evo}}$  was estimated by Powell. Because of restrictions on memory use only alignments referenced by the mouse were used for estimating  $E_{\text{evo}}$ .

## IMPLEMENTATION

The EHMM framework has been implemented in C. The HMM architecture and the evolutionary model are specified by a simple format. Gene prediction in a part of the human/mouse test set consisting of 174 alignments with a total length of 959,577 nucleotides takes 294 CPU seconds on a 1000 MHz, 500 Mb ram, AMD Athlon(tm) desktop computer.

## RESULTS

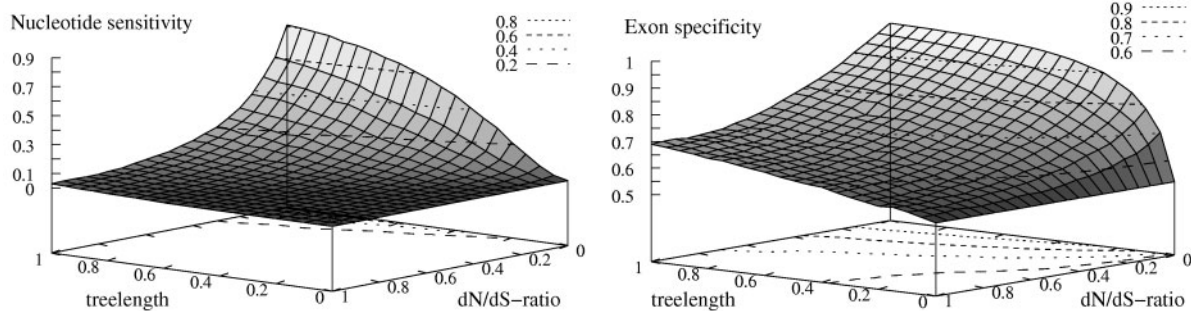
The benefits of modelling evolution with an EHMM approach are investigated through a number of simulation studies, and by a data set of orthologous mouse/human gene pairs. The performance is measured using the set of accuracy statistics conventionally used for gene-finding (Bursat and Guigo, 1996).

## Simulations

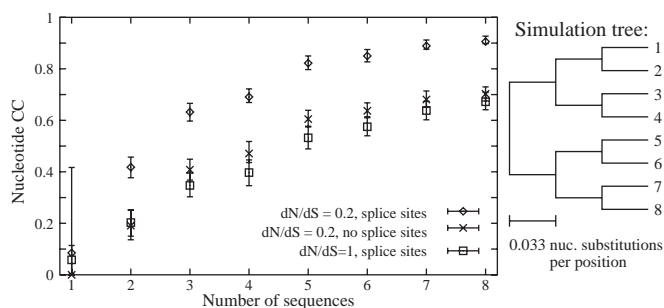
The benefit of modelling the evolutionary process for gene-finding will depend on the divergence between the sequences compared, and the difference in evolutionary pattern between the functional regions. In the EHMM sequence divergence is modeled by the overall substitution rate of each  $E_k$  and scaled by  $T$ . A key parameter for modelling the difference in evolutionary pattern between exons and introns is the  $dN/dS$ -ratio. The effect of these parameters on gene-finding performance was investigated by simulating a pair of aligned genomes under a range of parameter settings followed by gene finding using these true parameters. All parameters not under investigation were set at equal values between the states. The results are plotted in Figure 4.

The performance is seen to rise with both increasing tree length and falling  $dN/dS$ -ratio. At a  $dN/dS$ -ratio of 1 the performance remains nearly constant with tree length, due to the equal evolutionary patterns of introns and exons. At low  $dN/dS$ -ratios the performance increases fast with tree length, due to the pronounced difference in the evolutionary pattern of introns and exons. At zero tree length the performance remains constant with falling  $dN/dS$ -ratio, due to the lack of evolutionary events. The small rise in performance observable at a  $dN/dS$ -ratio of 1 is due to conservation of the splice site signals. As many positions are changed at long tree lengths it becomes increasingly likely that the splice sites are correctly predicted as belonging to the slowly evolving splice site states.

The effect of increasing the number of sequences in the alignment, and thereby the total amount of evolutionary information, was also tested by a simulation study. Eight genomes were simulated given a relating phylogenetic tree. Secondly the accuracy of gene predictions were found in alignments based on increasing subsets of these



**Fig. 4.** Simulation study illustrating the effect of dN/dS-ratio and tree length on prediction accuracy. Sensitivity at nucleotide level (left) and specificity at exon level (right) plotted against tree length and dN/dS-ratio. The unit of tree length is expected codon changes. The equilibrium distribution of the intron and coding states were set at uniformity. *ts/tv*-ratio is 1. Transition probabilities are set to make mean exon length = 50, mean intron length = 100. Total sequence length = 600 000.



**Fig. 5.** Simulation study illustrating the effect of aligning multiple genomes. Right part of figure: sensitivity at nucleotide level plotted against the number of sequences in an alignment, the bars mark the 95% confidence intervals. Left part of figure: the phylogenetic tree relating the sequences of the alignment. The sequences are included in increasing order. Two simulations were done with a dN/dS-ratio of 0.2, and one with a ratio of 1. One simulation was done with a model lacking splice sites between exons and introns. 100 iterations with sequence lengths of 100 000 were performed for each of the three parameter sets. All other parameters are set as in Figure 4.

genomes (see Figure 5).

The performance is seen to rise with the number of genomes compared. This rise can both be attributable to the diverging evolutionary pattern between alignment columns from introns and exons, or to the conservation of splice site signals. In order to illustrate both of these effects in isolation, two more runs of simulations were performed: One with the same dN/dS-ratio, but simulating exons which lack splice site signals, the other having splice sites, but with a dN/dS-ratio of 1 leaving the evolutionary pattern of introns and exons alike. Both the diverging evolutionary pattern and the conserved signals can be seen to contribute useful information (see Figure 5). Hence performance gains from modelling of both.

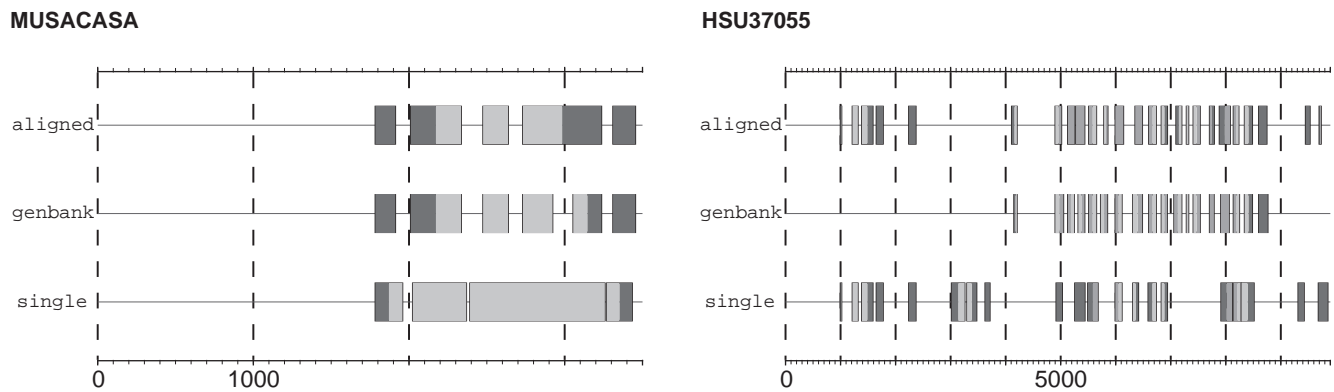
**Table 1.** Mean and range of the mouse/human parameter estimates from the evolutionary models of the intergenic/intron state and the inner codon state. The estimates of equilibrium frequencies are not shown. Both substitution rates ( $\mu$ ) are reported as the expected number of nucleotide changes per position, thus the estimated inner codon rate was divided by three to ease comparison

Evolutionary parameter values			
State	$\mu$	<i>ts/tv</i>	dN/dS
Intergenic/Intron	0.37 [0.35:0.38]	1.80 [1.65:1.89]	—
Inner codon	0.22 [0.21:0.23]	2.60 [2.50:2.82]	0.12 [0.09:0.14]

### Orthologous mouse and human gene pairs

For each of four disjunct subsets of the human/mouse test set a complete set of EHMM parameters were maximum likelihood estimated (see data sets). The parameter estimates of the evolutionary models show a distinct difference between the intergenic/intron state and the codon state (see Table 1). The estimates of the evolutionary rate of the intergenic/intron state would probably be higher, if promoter regions and other conserved signals were modeled by separate states. The estimates of the dN/dS ratios are in good accordance with those reported by Makalowski and Boguski (1998), dN/dS = 0.19, Nekrutenko *et al.* (2001), dN/dS = 0.11. The difference between the evolutionary parameter estimates indicate a strong potential gain for evolutionary-based gene finders.

For each of the four estimated parameter sets, the gene prediction accuracy was evaluated on the part of the test set not used in the estimation procedure. These evaluations were performed both on single and aligned sequences in order to quantify the effect of modelling the evolutionary process. When the EHMM is used on single sequences, which lack evolutionary information, each state's emission distribution is reduced to the equilibrium distribution of its



**Fig. 6.** Alignment (top rows) and single-sequence-based predictions (bottom rows) for the mouse gene MUSACASA (left) and the human gene HSU37055 (right). The GenBank annotation is shown in between the two predictions. Each exon is colored according to its own and the next exon's reading frame. These plots were produced using the visualization program gff2ps (Abril and Guigo, 2000).

**Table 2.** Accuracy statistics for gene predictions in the human/mouse data set. The statistics for ROSETTA (Batzoglou *et al.*, 2000) and GENSCAN (Burge and Karlin, 1997) are from Batzoglou *et al.* (2000), who did not report the correlation coefficient or overall exon statistics. Only the mean of the four separate evaluations are shown

Accuracy table					
Method	Nuc. Sn	Nuc. Sp	CC	Exon Sn.	Exon Sp
EHMM single	0.78	0.52	0.53	0.09	0.07
EHMM aligned	0.88	0.72	0.75	0.41	0.27
ROSETTA	0.95	0.97	—	—	—
GENSCAN	0.98	0.89	—	—	—

evolutionary model, the EHMM thus becoming a normal HMM.

The accuracy statistics of all four evaluations show the same pattern, the mean of which can be seen in Table 2. The performance on the aligned sequences is consistently better than on the single sequences. On the nucleotide level the biggest gain is in specificity. Exon level accuracy is very low for the single sequences, but experiences a *circa* 4-fold increase on the alignments.

The predictions on the single and the aligned sequences were also compared graphically, two representative examples are shown in Figure 6. Predictions based on the single sequences show a higher tendency to over-predict short exons, and skip short introns. Many of these mistakes are avoided on the aligned sequences, many of the exon boundaries are, however, still slightly wrong.

The splice site finders of the model are extremely simple: giving the two consensus nucleotides a high likelihood when slowly evolving. False splice sites with a high likelihood will therefore be common. It thus becomes

cheap to use false splice sites to enter and exit exons, explaining the short exons mis-predicted on the single sequences, and the wrong exon boundaries in both sets of predictions. The diverse evolutionary pattern of exons and introns obviously aid their discrimination on the aligned sequences. This effect is clearly seen in Figure 6, where short introns and exons are skipped on the single sequence but detected on the aligned. It is also registered in the improvement of the exon statistics (see Table 2).

Both the single and the alignment-based predictions predict a group of exons upstream of the true HSU27055 gene (see Figure 6). These are likely to be caused by a pseudo gene or a transposon, as no attempt was made to mask them out.

The experiments performed here show that gene finding can benefit from an EHMM approach when homologous sequences are available. The simple model used here is, however, not competitive with the state of the art methods (see Table 2). But any HMM-based gene finder, as GENSCAN, can be extended into an EHMM and thereby use the available evolutionary information to raise performance.

## DISCUSSION

The simulations demonstrated that the evolutionary pattern conveys useful information for gene finding. The performance of the EHMM on real data showed a pronounced improvement from single to aligned sequences. The absolute level of the accuracy statistics was, however, lower than those for leading single-sequence methods, as GENSCAN, or a comparative method as ROSETTA (see Table 2). This is not surprising as a minimal EHMM of eukaryotic gene structure was used. Obvious extensions to this model are: more advanced

splice site finders, models of ribosome binding site and promoter regions, non-geometric length distributions of exons, emission models of higher order (see below), and parameter estimations using larger data sets. Many of the successful non-comparative gene finders are HMM-based, e.g. GENSCAN and HMMgene (Krogh, 1997). Reimplementing one of these in the EHMM framework would result in a model with the same performance on single sequences, but with improved performance when homologous sequences are available. The existing comparative methods, such as ROSETTA or TWINSCAN, are limited to comparison of sequences from two genomes. The tree-based evolutionary analysis of an EHMM makes it capable of extracting information from additional genomes (see Figure 5).

Most HMM-based gene finders condition the emission probability of a given sequence position by its preceding positions, these are called higher-order models. This concept cannot be transferred to EHMMs without the development of 'conditional evolutionary models'. An approximation, which we call pseudo higher-order EHMM, is to make only the equilibrium frequencies of the evolutionary models dependent on the previous positions of the reference sequence. A pseudo higher-order EHMM will be equivalent to a higher-order HMM when used on the reference sequence alone, as the equilibrium frequencies will equal the emission distributions for single sequences. This property is attractive when extending existing HMMs to EHMMs.

The existing EHMM framework does not model gaps, but treats them as missing data. This strategy is not optimal for the process of gene finding, as gaps convey much information about the gene structure. For example, gaps in coding regions will occur in triplets and with low frequency. The optimal solution is to use an evolutionary model which also models gap structure, such as the TKF-model (Thorne *et al.*, 1991). The TKF-model is, however, computationally expensive, and its dependencies among columns make it inapplicable for the EHMM setting. Instead a simple two-state model of gap evolution is currently being implemented.

The optimal data for EHMM-based gene finding is a multiple alignment of full-length genomes. Such data is presently unavailable, due to the lack of multiple closely related full-length genomes and difficulties in constructing co-linear genomic maps. Much non-contiguous data from many different organisms is presently available in databases of genomic DNA, cDNA, and ESTs. A better strategy, at present, is therefore to search the reference genome against these databases, constructing a partial pairwise alignment for each organism. This alignment method resembles the strategy used by Korf *et al.* (2001) for constructing a 'conservation sequence' between mouse and human. The set of pairwise alignments can be

combined into a multiple alignment using their common gap-less reference sequence.

A challenge in constructing the alignment is not to introduce more noise than signal. Regions with wrongly aligned fragments will get a disturbed evolutionary pattern, lowering the likelihood of correct prediction. This risk can be reduced by only including high-scoring matches in the pairwise alignments.

A different strategy, as mentioned in the introduction, is to align and annotate in one combined step. The methods following this approach (Pachter *et al.*, 2001; Bafna and Huson, 2000; Blayo *et al.*, 1999; Scharling, 2001) exploit the characteristic evolutionary pattern of different functional regions in the alignment process, thereby reducing the risk of misalignment. The drawbacks are an assumption of co-linear genomes, a time complexity which is quadratic in genome length, and use of parameters which are not readily extended to more sequences.

The EHMM can be used for doing alignment by extending the idea of pair HMMs (Durbin *et al.*, 1998) to multiple sequences. The problem of parameterization is readily solved by its inherent use of evolutionary models. Such an extension will, however, still suffer from an exponential growth in time complexity with genome number. A heuristic linear time solution is to align only in a band of the alignment space. This band could be defined by proximity to a multiple alignment, as constructed above.

The combination of evidence from different sources is a well-known strategy for improving gene-finding performance (e.g. Staden and McLachlan, 1982; Gelfand *et al.*, 1996; Krogh, 2000). Information about the evolutionary process itself has until recently not been exploited. The EHMM approach suggests itself as a stringent model for incorporating this growing information source. The idea of combining structural and evolutionary analysis by the EHMM (Goldman *et al.*, 1996) should be useful in many other areas of biological sequence analysis.

## ACKNOWLEDGEMENTS

We thank Anders Krogh and Thomas Schou Larsen for patiently explaining the application of HMMs to gene finding and for early discussions of the model. Ziheng Yang is thanked for providing C-code for matrix diagonalization. Martin Knudsen is thanked for his help with the implementation. Tejs Scharling, Bjarne Knudsen, Christian Storm Pedersen, and Roald Forsberg are thanked for valuable discussions.

Mikkel Schierup and the anonymous referees are thanked for their helpful comments on the manuscript. The Research was supported in part by grant 21-00-0283 from the Danish Natural Science Research Council.

## SUPPLEMENTARY DATA

For Supplementary data, please refer to *Bioinformatics* online.

## REFERENCES

- Abril, J.F. and Guigo, R. (2000) gff2ps: visualizing genomic annotations. *Bioinformatics*, **16**, 743–744.
- Bafna, V. and Huson, D.H. (2000) The conserved exon method for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 3–12.
- Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B. and Lander, E.S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, **10**, 950–958.
- Blayo, P., Rouzé, P. and Sagot, M. (1999) Orphan gene finding—an exon assembly approach. Unpublished.
- Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
- Burset, M. and Guigo, R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.
- Churchill, G.A. (1989) Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.*, **51**, 79–94.
- Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.
- Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
- Gelfand, M.S., Mironov, A.A. and Pevzner, P.A. (1996) Gene recognition via spliced sequence alignment. *Proc. Natl Acad. Sci. USA*, **93**, 9061–9066.
- Goldman, N., Thorne, J.L. and Jones, D.T. (1996) Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.*, **263**, 196–208.
- Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 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.
- Knudsen, B. and Hein, J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**, 446–454.
- Korf, I., Flicek, P., Duan, D. and Brent, M.R. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**, S140–S148.
- Krogh, A. (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 179–186.
- Krogh, A. (1998) An introduction to Hidden Markov Models for biological sequences. In Salzberg, S.L., Searls, D.B. and Kasif, S. (eds), *Computational Methods in Molecular Biology*, Chapter 4, Elsevier, Amsterdam, pp. 45–63.
- Krogh, A. (2000) Using database matches with HMMGene for automated gene detection in *Drosophila*. *Genome Res.*, **10**, 523–528.
- Krogh, A., Mian, I.S. and Haussler, D. (1994) A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res.*, **22**, 4768–4778.
- Kulp, D., Haussler, D., Reese, M.G. and Eeckman, F.H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **4**, 134–142.
- Liò, P. and Goldman, N. (1998) Models of molecular evolution and phylogeny. *Genome Res.*, **8**, 1233–1244.
- Makalowski, W. and Boguski, M.S. (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA*, **95**, 9407–9412.
- Nekrutenko, A., Makova, K.D. and Li, W.-H. (2001) The  $k_A/k_S$  ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.*, **12**, 198–202.
- Pachter, L., Alexandersson, M. and Cawley, S. (2001) Applications of generalized pair hidden Markov models to alignment and gene finding problems. In Lengauer, T., Sankoff, D., Istrail, S., Pevzner, P. and Waterman, M. (eds), *Proceedings of the Fifth International Conference on Computational Biology (RECOMB)*, vol. 1, ACM Press, New York, pp. 241–248.
- Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992) *Numerical Recipes in C*, Second edn, Cambridge University Press, Cambridge.
- Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–285.
- Scharling, T. (2001) Master's thesis, (URL: <http://www.birc.dk/Publications/>), in Danish, *Gen-identifikation ved sekvens-sammenligning*, Aarhus University, Denmark.
- Staden, R. and McLachlan, A.D. (1982) Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res.*, **10**, 141–156.
- 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.
- Viterbi, A. (1967) Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Information Theory IT*, **13**, 260–269.
- Yang, Z. (2000) *Phylogenetic Analysis by Maximum Likelihood (PAML)*, 3rd edn, University College, London.