Gene Prediction with a Hidden Markov Model and a new Intron Submodel

Mario Stanke *and Stephan Waack†

July 17, 2003

Abstract

The problem of finding the genes in eukaryotic DNA sequences by computational methods is still not satisfactorily solved. Gene finding programs have achieved relatively high accuracy on short genomic sequences but do not perform well on longer sequences with an unknown number of genes in them. Here existing programs tend to predict many false exons. We have developed a new program, AUGUSTUS, for the ab initio prediction of protein coding genes in eukaryotic genomes. The program is based on a Hidden Markov Model and integrates a number of known methods and submodels. It employs a new way of modeling intron lengths. We use a new donor splice site model, a new model for a short region directly upstream of the donor splice site model that takes the reading frame into account and apply a method that allows better GC-content dependent parameter estimation. AUGUSTUS predicts on longer sequences far more human and drosophila genes accurately than the ab initio gene prediction programs we compared it with, while at the same time being more specific. A web interface for AUGUSTUS and the executable program are located at http://augustus.gobics.de. The datasets used for testing and training are available at
http://augustus.gobics.de/datasets/
Contact: mstanke@gwdg.de, waack@math.uni-goettingen.de

1 Introduction

Gene prediction programs typically use mathematical models of biological signals such as splice sites or the translation start and end points. With respect to what additional information they use, programs can be divided in two major groups. The so called *ab initio* programs use a training set with known gene structure for training the parameters of their models of the biological signals and the models for coding and non coding regions. Examples of ab initio programs are GENSCAN (Burge, 1997), GENIE (Kulp et al., 1996), HMMGene (Krogh, 1997) and GENEID (Parra et al., 2000).

*Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik,Universität Göttingen,Goldschmidtstraße 1 37077 Göttingen, Germany

†Institut für Numerische und Angewandte Mathematik,Universität Göttingen, Lotzestraße 16-18, 37083 Göttingen, Germany

Similarity-based programs use external information about the input sequence. Programs as GENEWISE (Birney and Durbin, 1997), PROCRUSTES (Gelfand et al., 1996) and GENOMESCAN (Yeh et al., 2001) make use of a homology to a known protein. Some programs instead use a second (syntenic) genomic sequence of another species. And exploit the similarities when both sequences code for similar proteins. Examples of programs of this type are AGenDA (Taher et al., 2003, Morgenstern et al., 2002), SGP-1 (Wiehe et al., 2001), TWINSCAN (Korf et al., 2001), DOUBLESCAN (Meyer and Durbin, 2002) and CEM (Bafna and Huson, 2000). For an overview of gene prediction methods see Mathé (2002). Similarity-based programs can often achieve a somewhat higher accuracy in gene prediction but require either that a similar enough protein or a second genomic sequence coding for similar proteins is known. A purely similarity-based program is stuck when there is no homology for the sequence under investigation. In that sense ab initio gene finding tools are more general applicable as they can be applied to sequences without known homologies.

The accuracy of ab initio gene finding tools is often evaluated on short genomic sequences containing exactly one gene and very little flanking DNA. In these short sequences the best programs perform fairly well. But as Guigó et al. (2000) pointed out, this is not a realistic setting for gene finding tools. The accuracy of GENSCAN, which is considered one of the best ab initio gene prediction programs for humans, drops significantly on the BAC-sized human sequences compiled by Guigó et al. The number of genes whose exact structure is correctly predicted by GENSCAN decreases from 40% to 18% . This loss of accuracy is due to the many false positive exons GENSCAN then finds in the intergenic region.

We have developed a new Hidden-Markov Model and implemented it in a program we call AUGUSTUS. AUGUSTUS uses a new method that allows a more accurate modeling of the intron lengths which could also be applied to other HMM-based gene prediction programs. Short introns typically have a length distribution clustering around a certain length. We model the length distribution of short introns very accurately and use a geometric distribution only for the lengths of long introns. The core of our splice site models is very simple as we use the empirical distribution as the probabilistic model. In case of the donor splice site, this empirical distribution is smoothed is a way that takes into account that patterns 'similar' to a frequent splice site pattern often also are frequent splice site patterns. We introduced a new model for bases -8 to -4 before the donor splice site which respects the reading frame of the exon. We applied a new method to train the model parameters dependent on the GC-content of the input sequence. AUGUSTUS performs much better than the other programs tested on two drosophila data sets. On a large sequence contig from the Adh region of drosophila melanogaster the number of exons *not* correctly predicted by AUGUSTUS is about half of the corresponding number for GENEID and GENIE. On a human data set with BAC-sized sequences AUGUSTUS predicts more than twice as many gene structures correctly than GENSCAN and GENEID

2 THE HIDDEN MARKOV MODEL UNDERLYING AUGUSTUS

A Hidden Markov Model is a probabilistic model. For the purpose of gene finding, it consists of states corresponding to a biological meaning (e.g. intron, exon, splice site) and allows transitions between these states in a biologically meaningful way (e.g. an acceptor splice site must follow an intron). The model defines a probability distribution on DNA sequences together with their gene structure. Programs based on these models often find a most likely gene structure given the DNA input sequence. For an introduction to Hidden Markov Models see for example Merkl and Waack (2002) or Durbin et al. (1999). Figure 1 shows the states of the Hidden Markov Model used in AUGUSTUS. Each state emits a random DNA string of possibly random length. The distribution of these strings as well as the transition probabilities between them were determined using a training set of annotated sequences for the respective species. In order to define this distribution for each state we made use of established models such as a Markov chain, a higher order windowed weight array model (WWAM) (Burge and Karlin, 1997), interpolated Markov Models (IMM) (Salzberg et al, 1997) and introduced a simple method we call similarity-based weighting of sequence patterns.

A *WWAM of order k and of window size* $2r+1$ is an inhomogeneous Markov Model of order k in which the probability of observing nucleotide x at position i given that the preceding k nucleotides are $x_1, \ldots, x_k$ is estimated by the relative frequency of observing x after nucleotides $x_1, \ldots, x_k$ in the training data at one of the positions in the window $i-r, \ldots, i+r$. For that purpose the training data is aligned with respect to the biological signal that is modeled.

By an *interpolated Markov Model of order* $k \geq 2$ we denote in this paper a Markov Model of order k for DNA sequences, in which for some sequence patterns $x_{i-k}, \ldots, x_{i-1}$, the probability of observing nucleotide x_i depends only on the $k-1$ preceding nucleotides instead of on all k preceding nucleotides, i.e. $p(x_i \mid x_{i-k}, \ldots, x_{i-1})$ is equal for all x_{i-k}. We here use a special case of the IMM described in (Salzberg et al 1997), in which only the transition probabilities of orders k and $k-1$ are considered and the respective interpolation weights are either 0 or 1. The conditional probability of observing nucleotide x_i after nucleotides $x_{i-k}, \ldots, x_{i-1}$ is

$$
\begin{aligned}
& p(x_i \mid x_{i-k}, \ldots, x_{i-1}) \\
= & \begin{cases} \frac{\#(x_{i-k}, \ldots, x_i)}{\#(x_{i-k}, \ldots, x_{i-1})} & \text{if } \#(x_{i-k}, \ldots, x_{i-1}) \geq 400; \\ \frac{\#(x_{i-k+1}, \ldots, x_i)}{\#(x_{i-k+1}, \ldots, x_{i-1})} & \text{otherwise.} \end{cases}
\end{aligned}
$$

Here the character $\#$ in front of a pattern denotes the frequency of the pattern in the training sequences (in the appropriate reading frame if applicable). We added a pseudo count of 5 to all pattern frequencies of sequence patterns with $k+1$ nucleotides.

In the following we describe the emission distribution for each state. Each state uses one or more simple submodels for different parts of the sequence. The parts are always considered independent. Figure 2 shows these parts and the underlying submodels for humans. The submodels are: **translation initiation motif:** WWAM of order 3 and window size 5 for the 20 bases before the translation start.

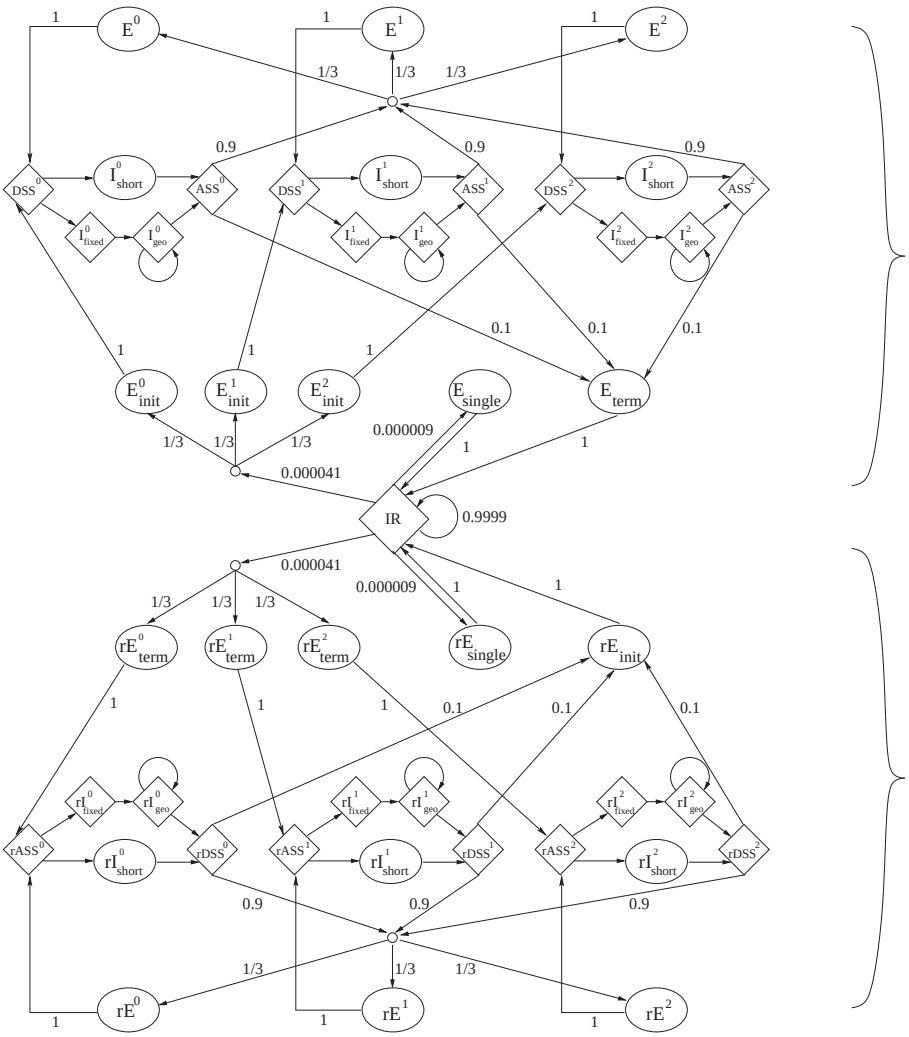

Figure 1: The states of AUGUSTUS and the possible transitions between them. The states with names beginning with r model the same as those without r but on the reverse strand. E_{single}: a single exon gene. E_{init}: The first coding exon of a multi exon gene (in this paper when we say exon we actually refer only to the coding part of the exons). DSS: the donor (5') splice site. I_{short}: an intron at most d nucleotides long. I_{fixed}: the first d nucleotides of a longer intron. I_{geo}: the individual nucleotides after the first d nucleotides of a longer intron. ASS: the acceptor (3') splice site including branch point. E: an internal (coding) exon. E_{term}: the last exon of a multi exon gene. IR: the 'intergenic region' between the genes modeled here. Diamonds stand for states which emit strings of fixed length, ovals for states with explicit length distribution. The numbers at the arrows are the transition probabilities. The remaining transition probabilities for the intron states are shown in Figure 5, they depend on the species. The exponents 0,1,2 stand for the phase of the reading frame. For an exon this is the position of the *last* coding nucleotide of the exon in its codon . For the other states the exponent stands for the phase of the preceding exon. The two small circles are silent states.

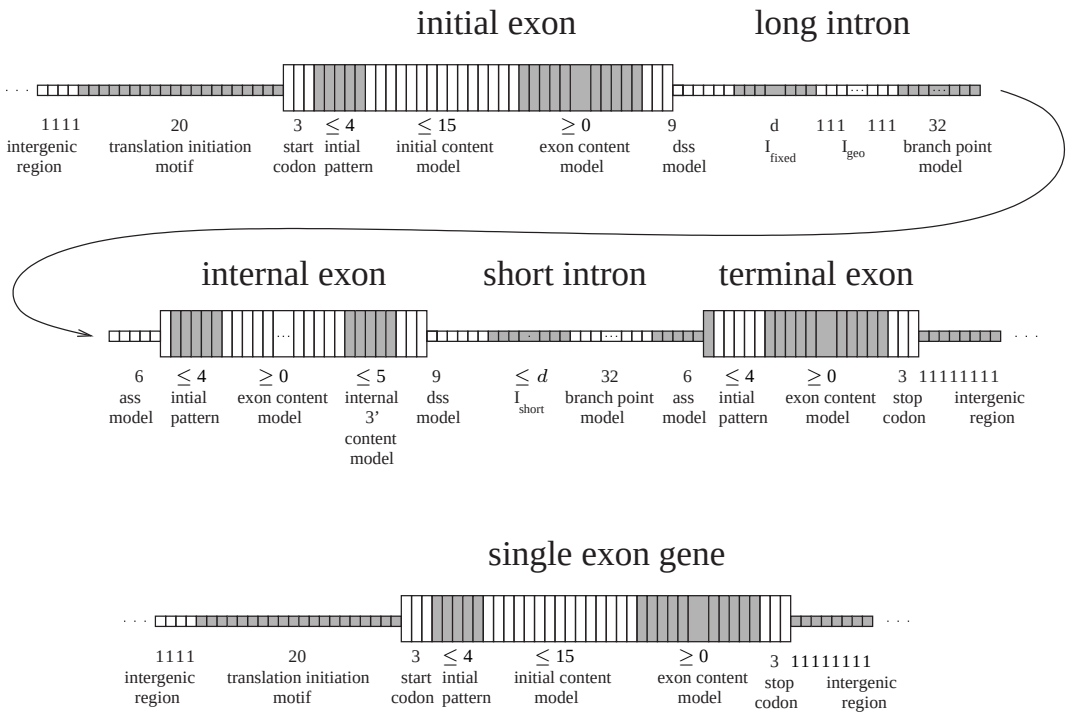

Figure 2: Example of a gene with 3 coding exons (above) and an intron-less gene (below). Certain parts of the DNA sequence are modeled using certain submodels. Below each part is written its length in the human version of the program and the name of its submodel. If an exon is shorter than the sum of the usual lengths of its submodels then the submodels are shorter or left out. The submodel most downstream are left out first.

start codon: Emit ATG with probability 1.
initial pattern: Emit pattern p of length at most 4 with the probability given by the relative frequency of this pattern in the corresponding reading frame among all coding sequences of the training set. The pattern has length 4 unless the exon length allows only shorter patterns. The reason for introducing this submodel is a technical one. If it was left out then the probability of the first bases after the start codon or after the acceptor splice site would be directly determined with a Markov model and therefore the nucleotides of the start codon or splice site would determine the emission probabilities of the following bases. But the start codon and splice site bases are always or often the same and an exception as far as typical coding sequences are concerned.
initial content model: interpolated 3-periodic Markov model of order 4. The length of the emitted sequence is 15 if the exon length allows it. The model is trained on the corresponding 15 nucleotides of single and initial exons of the training set. We also tried a corresponding terminal content model in the region around the stop codon as this was suggested by a bias in the distribution of these bases compared to the models

we actually use. But this model did not yield any improvement.
exon content model: interpolated Markov model of order 4 trained on all corresponding coding sequences of the training set. Only in earlier stages of AUGUSTUS with fewer submodels the order 5, which is more commonly used in other programs, yielded better results.
dss model: We only consider canonical splice sites obeying the GT-AG rule as this rule accounts for about 99% of mammalian splice sites (Burset 2000). The donor splice site model emits the 3 last nucleotides of the exon, then the consensus dinucleotide GT, and 4 more nucleotides of the intron (drosophila: 2 before, 4 after GT). For the distribution of the 7 free nucleotides we use a model we call *similarity-based sequence weighting*. The method of similarity-based weighting of sequence patterns is as follows. Given a fixed sequence pattern size, training patterns $q_1, \ldots, q_m$ and a similarity scoring function s, weighting pairs of patterns, we estimate the probability that a random pattern equals a given pattern q as

$$p(q) = c \sum_{i=1}^{m} s(q, q_i),$$

where c is chosen so that the sum of all $p(q)$ is 1. The choice of s depends on the particular purpose. For the donor splice site we use

$$s(r,q) = \begin{cases} 1 & \text{if } r = q; \\ 0.001 & \text{if } r \text{ and } q \text{ differ at exactly one pos.;} \\ 0 & \text{otherwise.} \end{cases}$$

This way, sequences obtained by a single point mutation from a typical splice site get a bonus in comparison with the empirical distribution. And the resulting distribution is the discretely smoothed empirical distribution which respects the complicated statistical dependencies that exist between the nucleotide positions.
I_{fixed}, I_{geo}, I_{short} **:** Markov models of order 4 trained on all non-coding sequences of the training set. I_{fixed} emits a sequence of exact length d, I_{geo} emits just one nucleotide at a time and I_{short} emits at most d nucleotides (see section about the intron length distribution).
branch point model: WWAM of order 3 and window size 7 emitting 32 nucleotides.
ass model: The acceptor splice site model emits 3 nucleotides of the intron before the AG dinucleotide consensus, then AG and the first nucleotide of the exon. A pattern of the 4 free nucleotides gets as probability the relative frequency of these 4 nucleotides at the corresponding positions in the training set.
internal 3' content model: interpolated 3-periodic Markov model of order 4 trained on the 5 nucleotides at positions -8 to -4 with respect to the donor splice site using all internal exons in the training set. Observe that this model, which helps locating the donor splice site, makes use of the reading frame of the coding nucleotides as opposed to the dss model for nucleotides -3,-2 and -1. This model is not used for drosophila.
stop codon: Emit TAG, TGA or TAA with probabilities 24%, 48% and 28%, respectively.
intergenic region: same as I_{geo}

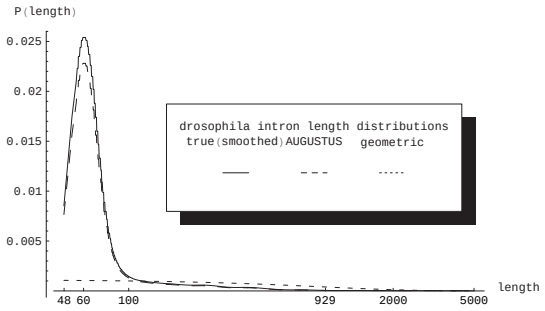

Figure 3: The smoothed length distribution of drosophila introns, the length distribution of introns of the AUGUSTUS model and the geometric distribution. The geometric distribution is a bad approximation for short introns. 63% of the introns are shorter than 100 nucleotides.

Intron Length Model

The geometric Approximation

Hidden Markov Models for gene prediction typically have one or more states modeling a biological intron. The states of such a model can have an explicit length distribution of the sequence emitted in this state or the length can be implicitly modeled by emitting just one nucleotide at a time but allowing to transition back to the same state. States with an explicit length distribution allow an accurate modeling of the length at the cost of computation time. If no further heuristic is used the computation time of the typical algorithms (Viterbi, forward algorithm) is at least proportional to the maximal possible length of this state. Introns can be very long: the human neurexin-3 gene on chromosome 14 has an intron of length 479 Kb (Wong et al., 2001). It is therefore practically infeasible to explicitly model the whole length distribution in a HMM. The method of using a state which emits just one nucleotide and allowing transitions back to the state is computationally efficient. The algorithms only require constant time for each position of the sequence for this state. But this option limits the length distribution of introns to a 'shifted' geometric distribution which assigns length $l > \delta$ the probability $q(1-q)^{l-1-\delta}$ with parameters $0 < q < 1$ and integer δ. δ would be the length of those parts of an intron which are modeled in other states as for example regions around the splice sites. For example the HMM based gene prediction programs GENSCAN, GENIE, TWINSCAN and DOUBLESCAN use a model in which the introns have a shifted geometric length distribution.

The solid line in Figure 3 and 4 shows the smoothed length distribution of drosophila introns in our training set of 320 genes. In both figures the horizontal axis is on logarithmic scale. Figure 4 also has the vertical axis on logarithmic scale so that the length distribution for large lengths can be visualized. The mean intron length is 896

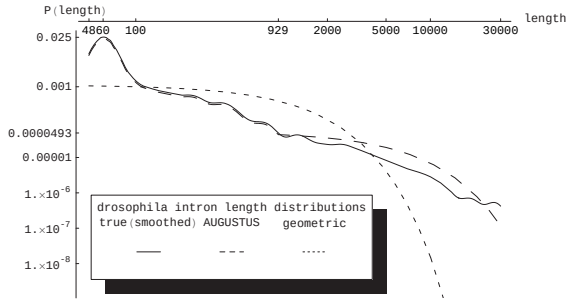

Figure 4: The same curves as in Figure 3 but with both axes on logarithmic scale. Up to $d = 929$ AUGUSTUS uses approximately the 'true' length distribution, the tail of AUGUSTUS' distribution is geometric, too. About 13% of the introns are longer than 929 nucleotides.

nucleotides. The figures also show the geometric length distribution with the parameter estimated by the maximum likelihood method: $P(L = l) = q(1 - q)^{l-1}$, with $q = 1/896$.
The graphs show two shortcomings of the geometric distribution as a model for intron lengths. One problem is that a (shifted) geometric distribution always assigns the highest probability to the shortest possible length. But in our drosophila test set the shortest intron had length 48 and there were 12 introns with a length between 48 and 52 but there were 223 introns with a length between 58 and 62. A program that uses the geometric intron distribution must either allow no such short introns or must assign a higher probability to their length than it assigns to any longer length. The other problem of a geometric distribution is that, when q is realistically chosen, long introns become much less likely than they really are. Reese et al. (2000) explain the fact that many long introns are not recognized by their program GENIE as follows "...the length distribution of introns, a geometric distribution that favors short introns, is the reason for so many split genes".

A new Way of modeling the Length Distribution

We combine states with and without explicit length distribution in order to model an *initial part* of length d of the length distribution more accurately and the remaining part with a geometric distribution. This makes the implicit length distribution much more accurate while at the same time not loosing too much of the computational efficiency. We use the model shown in Figure 5 for introns.

As intron we consider internally the part of the sequence between the donor splice site model and the branch point model, which is included in state ASS. Assume L is the length of a random intron and we know the distribution of L from the training

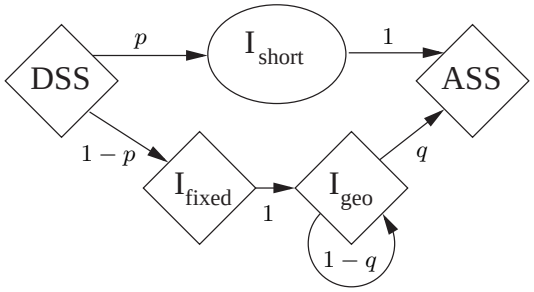

Figure 5: A proposition for an intron model. The arrows denote possible transitions and are marked with the transition probabilities.

data and let M be the random length of an intron generated by our model in Figure 5. The state I_{short} has an explicit length distribution with maximal length d, namely length l with $0 \leq l \leq d$ has probability $P(L = l)/P(L \leq l)$. The state I_{fixed} emits a string of fixed length d and the state I_{geo} emits just one nucleotide but implicitly has a geometric length distribution with parameter q. Each of the three intron states use the same 4th order Markov chain for emitting the nucleotides. There is a one-to-one correspondence between introns and paths from DSS to ASS. If the intron has length at most d the corresponding path goes through I_{short} and if it has length $l > d$ the path goes first through I_{fixed}, then $l - d$ times to state I_{geo} and then leaves I_{fixed} to ASS. The distribution of M is as follows. For $l \leq d$ we have $P(M = l) = pP(L = l)/P(L \leq l)$ (= transition probability to I_{short} times length probability). For $l > d$ we have $P(M = l) = (1 - p)(1 - q)^{l-d-1}q$ (= product of all transition probabilities). Now q, p and d are still free parameters. We set q such that the expectation of M, given $M > d$, is the expectation of L, given $L > d$, i.e. $d + 1/q = E[L \mid L > d]$. Then we set p such that $P(M = d + 1) = P(M = d)$ and there is no jump in the distribution of M between positions d and $d + 1$. Then it remains to choose the parameter d which is a tradeoff between accuracy (large d) and speed (small d). We choose d to be smallest such that $p \approx P(L \leq d)$. We get $q \approx 1/4894, p \approx 0.78, d = 929$ for drosophila and $q \approx 1/1688, p \approx 0.43, d = 584$ for humans. The running time of AUGUSTUS is about 6 minutes for the 1.6 megabases of the drosophila test set on a pc with 2.4 GHz.

This model architecture would also allow to use different content models for long and for short introns. Also additional splice site models could be integrated into the short intron model so that different splice site models could apply to short and long introns. This is suggested by the assumption that the splicing process is typically different for long and short introns (Lim and Burge, 2001). The resulting model would also allow to take dependencies between the donor and acceptor splice sites of short introns into account.

2.1 Algorithms and Options

AUGUSTUS computes for a given input DNA sequence the sequence of states and emission lengths that is most likely in this model, given the input sequence. In our model there is a one-to-one correspondence between gene structures one the one hand and sequences of states and their lengths on the other hand. So AUGUSTUS predicts the gene structure with the largest a-posteriori probability. It is found using the Viterbi-algorithm. For DNA input sequences with independent identically distributed nucleotides the expected running time grows linearly with the input sequence length. The memory requirement of AUGUSTUS also grows linearly with the sequence length. For long sequences AUGUSTUS needs roughly one mega byte per kilo base input sequence length. If the sequence is too long for the amount of memory available it is internally cut into pieces and it is assumed that the boundaries between the pieces lie in the intergenic region. The parameter specifying the length of the pieces was set to 400 Kb in the test cases below so that the program could be run on a pc with 512 MB RAM. AUGUSTUS can be run with options specifying whether it may predict partial genes (some exons missing in a gene at the boundaries of the input sequence), whether it may predict only complete genes, whether it must predict exactly one complete gene or whether is must predict at least one complete gene. The last two options mentioned are implemented by adding a second intergenic region state where each valid state path is forced to end in. The model specified in Figure 1 requires that genes must be separated by an intergenic region and also genes on opposite strand must not overlap. The idea of the "shadow" states for the simultaneous prediction of genes on both strands comes from Borodovsky et al (1993). But there are exceptions known to this rule. The human neurofibromatosis type I gene on chromosome 17 has three short genes on the opposite strand within one of its introns. Each of these internal genes has introns itself. For this reason AUGUSTUS has an option to ignore possible genes on the opposite strand. When this option is set, only the upper half of the states in Figure 1 is used for both the forward strand and its reverse complement. The prediction for both strands then consists of all the predictions for the forward and all the predictions for the reverse strand and the genes may overlap. In all the test runs mentioned below the options were set such that AUGUSTUS may predict partial genes and the prediction of overlapping genes is forbidden.

Results

We tested AUGUSTUS on four data sets which we call fly100, adh222, h178 and sag178.

fly100 is a set of 100 sequences of drosophila melanogaster with one gene each on the forward strand. 18 of the 100 genes were single exon genes. The mean sequence length is 16.1 kilo bases (shortest 2, longest 104 kilo bases). The sequences were retrieved from FlyBase and have been filtered for annotation errors and redundancies as described in section TRAINING.

adh222 is a single sequence of drosophila melanogaster and 2.9Mb long. It is a well-characterised sequence contig from the Adh region and has been used in the Genome

Annotation Assessment Project (GASP) (Reese, Hartzell et al., 2000). They constructed two sets of annotations. The first, smaller set, called std1, was chosen so that the genes in it are likely to be correctly annotated and the second larger set, called std3, was chosen to be as complete as possible "while maintaining some confidence" about the correctness. In the corrected version std1 contains 38 genes with a total of 111 exons and std3 contains 222 genes with a total of 909 exons. The genes lie on both strands. Both the authors of GENIE and of GENEID (Parra et al., 2000) have used these two annotation sets for testing their programs. It should be noted that std1 was chosen to contain only splice sites with a high score in a neural network model used in GENIE and provided by M.Reese.

h178 is a set of 178 human genomic sequences with one complete gene each. Each contains one gene and a little flanking DNA. The sequences are from EMBL, were compiled by Guigó et al (2000) and have also been used by the author of GENSCAN for evaluation (Yeh et al., 2001). The mean sequence length is 7169 bases (shortest 622, longest 86640 bases).

sag178 is a set of 43 sequences with 178 human genes on both strands. These sequences were also taken from Guigó et al (2000) and are semi-artificial in the following sense. Guigó et al took the 178 sequences from h178 and generated long intergenic regions randomly using a Markov model of order 5. They write "Some of the resulting parameters, such as average G+C content of 40%, a gene every 43Kb, and a coding density of 2.3% are in agreement with that for the overall human genome." 40 of the 178 genes were single exon genes. The mean sequence length is 177 kilo bases (shortest 70, longest 282 kilo bases) and each sequence contained on the average 4.1 genes.

We measured the gene prediction accuracy with the usual measures, sensitivity and specificity. For a feature (coding base, exon, gene) the sensitivity is defined as the number of correctly predicted features divided by the number of annotated features. The specificity is the number of correctly predicted features divided by the number of predicted features. A predicted exon is considered correct if both splice sites are at the annotated position of an exon. A predicted gene is considered correct if all exons are correctly predicted and no additional exons not in the annotation. Predicted partial genes were counted as predicted genes. For each data set these measures were computed globally (once for all sequences together) and in sag178 and adh222 the forward and backward strands were treated as different sequences.

Comparison to other Programs

For comparison we used GENSCAN (version 1.0), GENEID and GENIE. We took GENSCAN as it is the most commonly used gene prediction program and as it is considered one of the best programs for humans. Also our HMM is similar to that of GENSCAN. GENSCAN was run using its human parameter set for both human and drosophila as recommended. We used GENEID (version 1.1) as there is a special drosophila parameter set available for it and as it uses a different approach not modeling the lengths. GENEID first finds splice site candidates, then exon candidates using the splice site candidates and then genes using the exon candidates. GENEID was run using the parameter sets human3iso.param and dros.param, respectively. In one case we also compare to GENIE, because this program compared favorably to the other ab

initio programs in the GASP experiment. GENSCAN was not run on the Adh region as it required too much memory. GENSCAN and GENEID were downloaded from the Internet.

Tables 1 to 4 show a summary of the results of the programs on the test sets. On the drosophila data sets (Tables 1 and 2) AUGUSTUS outperforms the two other programs on each of the three levels. On data set fly100 it predicts 52% of the genes correct, GENSCAN and GENEID only 37% and 31%, respectively. More than 3 out of 4 exons predicted by GENSCAN are false. On the human data set h178 with short single gene sequences (Table 3) AUGUSTUS and GENSCAN are similarly accurate with respect to the mean of sensitivity and specificity on the base and exon level. GENSCAN is more sensitive, AUGUSTUS more specific. GENEID is worse here. AUGUSTUS predicts more GENES (82) correctly than GENSCAN (71) and GENEID (24). On the long sequences in sag178 containing the same genes (Figure 4) AUGUSTUS predicts still 40% of the annotated genes exactly correct, GENSCAN and GENEID only 18% and 17%. GENSCAN here often 'adds' short exons to an annotated gene and is therefore much less specific than GENEID and AUGUSTUS.

Comparison to Variants of AUGUSTUS

In order to find out to which extend the new methods or submodels contribute to the accuracy of AUGUSTUS, we compared AUGUSTUS to versions of AUGUSTUS where one or more feature (method or submodel) was changed. We did this separately for human and drosophila but summarized the results for the two datasets for each species. In particular we use as an – admittedly – coarse measure the mean increase in sensitivity and specificity on the exon and gene level when the feature is used as compared to when the feature left out. For example, let $\Delta\text{sn}^i_{\text{exon}}$ be the difference between the sensitivity on the exon level on dataset $i \in \{1, 2\}$ of AUGUSTUS and AUGUSTUS with some feature changed. We weighted the two datasets for each species with the number of annotated genes n_1 and n_2 in the two datasets used to determine the accuracy measure, here the sensitivity. Then $\Delta\text{sn}_{\text{exon}} = (n_1 \cdot \Delta\text{sn}^1_{\text{exon}} + n_2 \cdot \Delta\text{sn}^2_{\text{exon}})/(n_1 + n_2)$ denotes the mean increase in exon sensitivity. We use

$$r := (\Delta\text{sn}_{\text{exon}} + \Delta\text{sp}_{\text{exon}} + \Delta\text{sn}_{\text{gene}} + \Delta\text{sp}_{\text{gene}})/4$$

as a measure to give the reader an idea of the relevance of the feature of the model. Table 5 shows for a selected number of features the relative improvement r. The first line refers to a version of AUGUSTUS, where the intron length was modeled using a shifted geometric distribution with minimum length 48 and the parameter estimated with the maximum likelihood method. The second line refers to the version of AUGUSTUS, where only the initial pattern model was left out, i.e. the start codon model or the ASS model is directly followed by a Markov content model. The third line refers to the version of AUGUSTUS where the donor splice site model simply uses the empirical distribution of the patterns (with pseudo counts). The fourth line refers to the version where all IHMMs were substituted by HMMs of the same order. This mostly effects the internal 3' content model in the human version. The fifth line refers to the version of AUGUSTUS where the internal 3' content model was left out. The last line refers to

AUGUSTUS where all the above changes are made. The largest improvement through a single new feature is obtained for drosophila with the introduction of the new intron length model.

We examined whether the improvement in exon sensitivity for drosophila by introducing the new intron length model might by explained by simple chance. For each of the exons of the 138 genes used to calculate the two exon sensitivities for drosophila we observe two dependent Bernoulli-random variables determining whether it was correctly predicted in the two runs, with or without the feature. The McNemar test for dependent samples yielded a p-value of 0.000034, so that an improvement simply by chance can be ruled out in this case.

2.2 Discussion

The reason that the new intron model does not improve much the predictions for humans can be explained by the fact that short human introns have a much less characteristic length than short drosophila lengths. The methods and models examined in Table 5 do not fully expain the increase in accuracy of AUGUSTUS in comparison to the other programs even when taking into consideration that the combination of several changes may yield a better improvement than expected from the inprovements that the individual changes yield alone. We do not know a single new method or idea that may explain this improvement. The fact that content models of order 4 yield better accuracy results than those of order 5 might be astonishing, as there are enough training data for training models of order 5 and models of higher order model the real distribution more accurate than models of lower order. We conjecture the following explanation for it. In theory a perfect program should consider the biological signals for prediction instead of statistical features of the coding and non-coding sequences because – probably – most of the sequence has no function for the transcription and translation process. For current state-of-the-art programs taking these statistical features into account by using content models helps improving accuracy. But not rarely the wrong content model yields a higher probability for a strech of sequence than the correct one, e.g. an untypical short exon or a stretch of non-coding sequence that gets a high probability in an exon model. Our observation is that the more 'accurate' the content models are, the larger are the differences in the probabilities that a stretch of sequence gets in the different content models. This means that the 'decisions' are made more by the content models than by the signal models and errors of the content models have a lower chance of being corrected by the signal models.

2.3 Future plans

AUGUSTUS performs significantly better on long sequences than other ab initio methods. But still about 20% of the exons are not predicted exactly and 7%-10% are missed completely. Thus it seems unnatural to dispense with extrinsic information about a DNA sequence in cases where such information indeed is available. We plan to plausibly integrate into the model information from database searches (both EST and Protein) and evidence about functional parts of the sequence using DIALIGN 2 (Morgenstern, 1999) for syntenic sequences.

fly100		AUGUSTUS	GENSCAN	GENEID
base	sn	97	97	95
	sp	59	33	53
exon	sn	80	68	65
	sp	49	22	39
gene	sn	52	37	31
	sp	27	10	14

Table 1: Accuracy results on drosophila data set fly100. Only genes on the forward strand were considered. A part of the 'false' positives accounting for the low specificity of all methods probably can be attributed to un-annotated genes in the sequences.

adh222		AUGUSTUS	GENEID	GENIE
base	sn*	98	96	96
	sp*	93	92	92
exon	sn*	86	71	70
	sp*	66	62	57
gene	sn*	71	47	40
	sp*	39	33	29

Table 2: Accuracy results on drosophila data set adh222. The asterisk (*) denotes that sensitivity and specificity were measured using two different sets of annotations. The sensitivity refers to std1 and the specificity refers to std3. The values for GENIE are taken from Reese, Hartzell et al. (2000).

h178		AUGUSTUS	GENSCAN	GENEID
base	sn	93	97	89
	sp	90	86	91
exon	sn	80	83	66
	sp	80	75	75
gene	sn	46	40	14
	sp	45	36	13

Table 3: Accuracy results on human data sets h178.

sag178		AUGUSTUS	GENSCAN	GENEID
base	sn	93	94	89
	sp	81	64	78
exon	sn	78	68	67
	sp	71	45	60
gene	sn	40	18	17
	sp	35	14	17

Table 4: Accuracy results on human data set sag178. AUGUSTUS predicts 75 of the 178 genes exactly correct, GENSCAN and GENEID predict only 32 and 31 genes correct, respectively. The gene level accuracy measures of GENSCAN on these long genomic sequences are similar to those reported in Korf et al. (2001) for long mouse sequences with mean length 112 Kb (sensitivity: 15-17, specificity: 11-16).

feature	human	fly
intron length model	0.3	3.4
initial pattern	1.6	1.0
similarity-based weighting	1.0	1.0
IHMM	1.8	0.0
internal 3' content model	0.8	n.a.
all of the above	4.7	6.8

Table 5: The relative mean improvement of sensitivity and specificity on exon and gene level caused by different features of the program. The largest increase in accuracy through a single feature is attributed to the new intron length model, but only for drosophila.

Training

Data Sets

The training set for the human version of the program was retrieved in October 2002 from Genbank. Sequences with inconsistent notation were deleted as well as sequences that were overlapping with a sequence in one of the human test sets. This was done using BLASTN with an E-value cutoff of 1e-100 (blastall -p blastn ... -S 1 -G 9 -q -9 -e 1e-100). The rest was cleaned for redundancies and 1284 sequences with one gene each remained. Additionally we use for the human parameter set the splice sites from 11739 human introns that were each not contained in the test sets (data originally retrieved from http://genomic.sanger.ac.uk/spldb/HumanCanonicalSites.ESTsupp). For the drosophila training and test set we took single gene sequences from FlyBase in December 2001. These were cleaned for genes with known alternative splicing, incomplete annotation, in-frame stop codons, non-canonical splice sites and for redundancies within the data set. The resulting 420 sequences were randomly divided into a training set of 320 and a test set of 100 sequences again with no BLAST match with E-value smaller than 1e-100 between the data sets. For the runs on the adh222 test set, we took these 420 drosophila sequences and removed those 20 sequences that had a BLASTN hit with E-value less than 10^{-10} when run on the Adh sequence. Except for the different training sets, the same parameters were used for training and testing the two test sets for each species. For comparison: The training set of GENSCAN consists of 380 single gene sequences plus additional unpublished 1619 cDNA sequences. Comparing the genomic sequences of h178 and the 380 training genes of GENSCAN shows that 91 of the 178 test genes have a blast hit in GENSCANs training set with E-value below 1e-100, most of them because the same gene is annotated. The data sets used here can be downloaded from http://augustus.gobics.de/datasets/.

2.4 Taking GC-content into Account

As Burge has pointed out in his thesis, sequence composition strongly correlates with GC content of the sequence. Burge took this into account by using 4 classes of GC content ($< 43\%, 43\% - 51\%, 51\% - 57\%, > 57\%$) and building 4 different parameter

sets by estimating most of the parameters only from sequences of the class. For a given input sequence the model parameters of its GC class are more appropriate but were estimated using only about one fourth of the training sequences. But for a reliable estimation of the parameters a large enough training set is important.
We use a method that allows us to use even more GC content specific parameter sets without the disadvantage of decreasing the training set size. We generate 10 different parameter sets for different GC contents of the input sequence. But for constructing each of these parameter sets we use *all* sequences of the training set. In constructing a parameter set for mean GC content α we weighed each sequence of the training set with an integer weight $1 \leq w \leq 10$ depending on its GC content β. Similar GC contents get a higher weight. As a weight function we used

$$w(\alpha, \beta) = \lceil 10 \exp(-200(\alpha - \beta)^2) \rceil \quad , (0 < \alpha, \beta < 1)$$

($\lceil \ldots \rceil$ means rounding up). We then trained the parameters as if each sequence was w times in the training set (Burges method can be regarded as a special case in which $w(\alpha, \beta) = 1$ if α and β are in the same class and $w = 0$, otherwise.). For the prediction AUGUSTUS chooses for each input sequence the parameter set with mean GC content closest to the one of the input sequence. The length distributions, transition probabilities and splice site models were trained independently of GC content. For all other submodels there are 10 different versions.

2.5 Smoothing Lengths

The length distributions of the introns and the 4 exon types were received by smoothing the empirical lengths with a normal distribution as kernel function and with variable bandwidth. I.e. if $l_1, \ldots, l_n$ (possibly with repeats) were the observed lengths of some state type in the training set then we used the length distribution

$$P(L = l) = \frac{1}{n} \sum_{i=1}^{n} \varphi_{l_i}(l_i - l)$$

for that state type. Here, φ_k is for $k = 1, 2, \ldots$ a discrete version of the normal distribution with mean 0 and standard deviation $c \cdot k$. The constant c was chosen to avoid over-smoothing on the one hand and a 'skyline effect' on the other hand.

REFERENCES

Bafna, V. and Huson, D. (2000) The conserved exon method for gene finding. *Bioinformatics*, **16**, 190-202.

Birney, E. and Durbin, R. (1997) Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison. *Ismb*, **5**, 56-64.

Borodovsky, M. and McIninch, J. (1993) GENMARK: parallel gene recognition for both DNA strands. *Comp.Chem.* **17**, 123-133.

Burge, C.B. (1997) Identification of Genes in Human Genomic DNA. PH.D. thesis.

Burge C.B. and Karlin S. (1997) Prediction of Complete Gene Structures in Human Gemonic DNA. *Journal of Molecular Biology* **268**, 78-94.

Burset, M., Seledtsov, I.A. and Solovyev, V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364-4375.

Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. (1999) *Biological Sequence Analysis*. Cambridge Univ Press.

Gelfand, M.S., Mironov, A.A. and Pevzner, P.A. (1996). Gene recognition via spliced alignment. *PNAS*, **93**, 9061-9066.

Guigó, R., Agarwal P., Abril J., Burset, M. and Fickett, J.W. (2000) An Assessment of Gene Prediction Accuracy in Large DNA Sequences. *Genome Res.*, **10**, 1631-1642.

Korf, I., Flicek, P., Duan, D. and Brent, M.R. (2001) Integrating Genomic Homology into Gene Structure Prediction. *Bioinformatics*, **1**, S1-S9.

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.

Kulp, D., Haussler, D., Reese, M.G. and Eeckmann, F.H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Intell. Systems Mol. Biol.*, **4**, 134-142.

Lim, L.P. and Burge, C.B. (2001) A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci*, **98**,11193-11198.

Mathé, C., Sagot, M.-F., Schiex, T. and Rouzé, P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, **30**, 4103-4117.

Merkl R. and Waack S. (2002) *Bioinformatik Interaktiv*. Wiley-VCH.

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

Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211-218.

Morgenstern, B., Rinner, O., Abdeddaïm, S., Haase, D., Mayer, K., Dress, A. and Mewes, H.-W. (2002) Exon Discovery by Genomic Sequence Alignment. *Bioinformatics*, **18**, 777-787.

Parra, G., Blanco, E. and Guigó, R. (2000) GeneID in Drosophila. *Genome Research*, **10**, 391-393.

Reese, M.G. (2000) Computational prediction of gene structure and regulation in the genome of Drosophila melanogaster. PH.D. thesis.

Reese, M.G., Kulp, D., Tammana, H. and Haussler, D. (2000) Genie – Gene finding in Drosophila Melanogaster *Genome Research*, **10**, 529-538.

Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F. and Lewis, S. (2000) Genome Annotation Assessment in Drosophila melanogaster *Genome Research*, **10**, 483-501.

Taher, L., Rinner, O., Gargh, S., Sczyrba, A., Brudno, M., Batzoglou, S. and Morgenstern, M. (2003) Homology-based gene prediction. *Bioinformatics*, **19**, in press.

Wiehe, T. ,Gebauer-Jung, S., MItchell-Olds, T. and Guigó, R. (2001) SGP-1: Prediction and Validation of Homologous Genes Based on Sequence Alignments. *Genome Research*, **11**,1574-1583.

Wong, G.K.-S., Passey, D.A. and Yu, J. (2001) Most of the Human Genome is Transcribed, *Genome Research*, **11**, 1975-1977.

Yeh, R.F., Lim, L.P. and Burge, C.B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.*, **11**, 803-816.