Chapter 3

Ab initio prediction of mouse and human genes with DOUBLESCAN

3.1 Introduction and motivation

The architecture of the pair HMM underlying DOUBLESCAN as described in Chapter 2 is suited for the comparative prediction of pairs of genes in any pair of related eukaryotic organisms. Only by setting up the pair HMM's transition and emission probabilities with a training set of known pairs of genes, is DOUBLESCAN specialised for a certain purpose.

In this chapter, we show that DOUBLESCAN can be used to analyse pairs of mouse and human DNA sequences. For this, DOUBLESCAN'S emission probabilities were derived and its transition probabilities optimised with a training set of known mouse and human gene pairs [JBD99], *see* Section 2.3 for the derivation of parameters and Section A.1 in Appendix A for a description of the training set. Once the parameters of the pair HMM have been set up, DOUBLESCAN is applied to a test set of 80 pairs of known orthologous mouse and human genes [Pac99], *see* Section A.2 in Appendix A. Note that the only input information to DOUBLESCAN are the letters of the two DNA sequences. In particular, the sequences are not masked for repeats.

This chapter evaluates the performance of DOUBLESCAN for finding genes and compares it to the performance of GENSCAN [BK97], a non-comparative *ab initio* gene prediction method. We then briefly discuss the alignments of the DNA sequences generated during the gene prediction and conclude the chapter by showing that the Stepping Stone algorithm provides a good solution for the analysis of long DNA sequences by comparing its predicted state paths

and gene predictions to the optimal ones derived with DOUBLESCAN using the Hirschberg algorithm.

3.2 Results

The results of the comparison are shown in Table 3.1. In the following, the term prediction refers to the results retrieved by DOUBLESCAN or GENSCAN, whereas the term annotation refers to the gene structures in the data base from which the genes of the training and test sets have been derived.

GENSCAN is a non-comparative ab initio gene prediction method which employs an explicit state duration HMM. It is capable of predicting partial, complete and multiple genes. Its HMM contains separate states for the exon of a single exon gene and for initial and terminal exons, as well as states for promoter, 5' untranslated region, 3' untranslated region and the poly-A signal. It uses different parameter sets according to the GC contents of the input DNA sequence.

The first thing to note is that the pair HMM with states for UTR-splicing improves the overall performance of DOUBLESCAN, especially the sensitivity and specificity for stop codons, the specificity for start codons and exons and the sensitivity and specificity for genes as well as the rate of wrong genes. 16 % of the overlapping genes are turned into correctly predicted genes and 42 % of the wrong genes are completely removed when including UTR-splicing into the model, while only 5 % of the correctly predicted genes are turned into just overlapping genes. Instead of a correctly predicted start codon, these overlapping genes have a splice site in close vicinity 5' to the annotated start codon which is not predicted or their initial exons are completely missing in the prediction. Given its superior performance, DOUBLESCAN including the states for UTR-splicing will be taken as the reference model for DOUBLESCAN. DOUBLESCAN including UTR-splicing still has a 14 % rate of wrong genes corresponding to 30 genes which are predicted in addition to those that overlap the annotated gene in each DNA sequence. 53 % of the wrong genes are short (less than 106 base pairs length) complete single exon genes, 13 % are complete two exon genes with a long intron (more than 656 base pairs length) and short coding length (less than 52 base pairs length), 7 % are complete two exon genes with a short intron (less than 17 base pairs length) and short coding length (less than 49 base pairs length) and the remaining 27 % are partial genes.

If we post-process DOUBLESCAN'S results as described in Section A.3 in Appendix A, all of

	DOUBLESCAN without UTR-splicing	DOUBLESCAN	DOUBLESCAN including post-processing	GENSCAN
Gene				
Sensitivity	0.51	0.57	0.57	0.47
Specificity	0.35	0.43	0.50	0.46
Genes overlapping	0.42	0.44	0.46	0.53
Genes missing	0	0	0.01	0
Genes wrong	0.23	0.14	0.04	0.01
Start Codon				
Sensitivity	0.77	0.78	0.75	0.73
Specificity	0.64	0.67	0.78	0.91
Stop Codon				
Sensitivity	0.86	0.91	0.89	0.88
Specificity	0.70	0.74	0.86	0.97
Exon				
Feature Level				
Sensitivity	0.79	0.81	0.80	0.84
Specificity	0.68	0.74	0.79	0.82
Exons overlapping	0.16	0.15	0.15	0.12
Exons missing	0.03	0.03	0.05	0.03
Exons wrong	0.16	0.10	0.06	0.06
Nucleotide Level				
Sensitivity	0.97	0.97	0.96	0.98
Specificity	0.97	0.98	0.99	0.94

Table 3.1: Performance figures for **DOUBLESCAN** without UTR-splicing, **DOUBLESCAN,** DOUBLESCAN including post-processing and **GENSCAN** on the test set. The predictions by DOUBLESCAN were generated using the Stepping Stone algorithm. Sensitivity is defined **as** the fraction of annotated features which are correctly predicted. Specificity is defined **as** the fraction of predicted features which match an annotated feature. For start and stop codons, sensitivity and specificity are shown at feature level, i.e. for entire codons. At feature level, sensitivity and specificity **as** well **as** the fraction of annotated exons which overlap a predicted exon (Exons overlapping), the fraction of annotated exons which do not overlap any predicted exon (Exons missing) and the fraction of predicted exons which do not overlap any annotated exon (Exons wrong) are given. At gene level, sensitivity and specificity are detailed **as** well **as** the fraction of annotated genes which overlap a predicted gene (Genes overlapping), the fraction of annotated genes which do not overlap any predicted gene (Genes missing) and the fraction of predicted genes which do not overlap any annotated gene (Genes wrong).

the wrong complete genes, corresponding to 73 % of the wrong genes, are removed. This post-processing steps also removes ten (10.7 %) of the overlapping genes. Six of them are complete short single exon genes which overlap an exon of the annotated gene. Two other overlapping genes which are removed in the post-processing step are complete two exon genes with a small coding length (less than 94 base pairs length) which have only a small overlap with the annotated genes. However, the post-processing step also removes two complete, overlapping multi-exon genes which overlap the annotated genes in most of their exons, but which each have one short intron (of 39 base pairs and 45 base pairs length, respectively) due to a mis-predicted exon. Overall, the post-processing step improves the performance considerably. It keeps the sensitivity at gene level unchanged while at the same time improving the specificity by 7 % and lowering the rate of wrong genes by 10 %. For start codons, it slightly lowers the sensitivity by 3 % while at the same time raising the specificity by 11 %. The same tendency is shown for stop codons where the sensitivity is lowered by 2 % while the specificity improves by 12 %. For exons, the performance at nucleotide level remains almost unchanged. At exon level, the sensitivity is lowered by 1 % while the specificity is increased by 5 %. Given the overall positive effect of the post-processing step, we discuss in the following parts of this chapter the results of DOUBLESCAN after post-processing unless otherwise stated.

Both for DOUBLESCAN and GENSCAN, the performance for stop codons is significantly higher than for start codons, the main reason being that in-frame start codons can be found both at the translation start as well as in frame within exons, while in-frame stop codons can only be found at the translation end. The sensitivity of DOUBLESCAN for start codons is 2 % higher than that of GENSCAN, but its specificity is 13 % lower than that of GENSCAN. DOUBLESCAN'S sensitivity for stop codons is slightly higher than that of GENSCAN, while its specificity is 11 % lower than that of GENSCAN. Unlike DOUBLESCAN, GENSCAN has dedicated states for a promoter and the 5' untranslated region which model the region 5' of the translation start. These extra states implement detailed knowledge about the upstream region of some genes and can therefore help to position the start codon correctly. In addition, GENSCAN is biased towards starting and finishing the predicted annotation within the intergenic state. Within GENSCAN, also the region 3' of the translation end has dedicated states which model the 3' untranslated region and a poly-A signal. However, without this extra information, DOUBLESCAN has a high sensitivity for both start and stop codons using only similarity information between the two DNA sequences.

DOUBLESCAN'S sensitivity for exons at nucleotide level is high, the sensitivity being 2 % lower and the specificity 5 % higher than those of GENSCAN. At exon level its sensitivity is 4 % and its specificity 3 % lower than GENSCAN. The difference in performance for exons between the nucleotide and exon level can be explained by cases in which two or more predicted genes overlap one annotated gene such that the overlap between the annotated and the predicted exons is large, but not perfect.

At gene level, DOUBLESCAN has a significantly higher sensitivity (10 %) and also higher specificity (4 %) than GENSCAN. Three of the gene pairs can not be predicted correctly by DOUBLESCAN as the configuration of annotated genes can not be modelled by the underlying pair HMM. One of the three gene gene pairs can not be modelled as the initial exons consist only of a start codon. The other two pairs of genes lie in pairs of sequences for which one sequence starts with intergenic subsequence 5' to the start codon and the other sequence starts directly with the start codon. Removing the corresponding three sequence pairs would improve the performance by up to 3 %. The 1 % rate of missing genes for DOUBLESCAN corresponds to one overlapping gene which is removed in the post-processing step.

In order to see whether or not DOUBLESCAN and GENSCAN preferentially detect different types of genes, we have compared the genes which were correctly predicted by one of the two methods to those predicted by the other method. About half (44%) of the genes which were found by DOUBLESCAN were incorrectly predicted by GENSCAN. Conversely, 32 % of the genes found by GENSCAN were not correctly predicted by DOUBLESCAN. By far the most common reason why a gene is correctly predicted by one method and incorrectly predicted by the other one is that the start codon is not found correctly or not found at all (accounting for 55 % of the genes found by DOUBLESCAN and not correctly predicted by GENSCAN, and for 58 % of the genes found by GENSCAN and not correctly predicted by DOUBLESCAN). The next common causes are incorrect splicing (accounting for 30 % of the genes found by DOUBLESCAN and not correctly predicted by GENSCAN, and for 21 % of the genes found by GENSCAN and not correctly predicted by DOUBLESCAN) and the wrong or missing prediction of the stop codon (accounting for 23 % of the genes found by DOUBLESCAN and not correctly predicted by GENSCAN, and for 25 % of the genes found by GENSCAN and not correctly predicted by DOUBLESCAN). Interestingly, GENSCAN tends to miss out whole terminal exons whereas DOUBLESCAN only gets the 3' end of the terminal exon wrong by introducing a 5' splice site in close vicinity 5' to the annotated stop codon. Overall, DOUBLESCAN and

GENSCAN complement each other, but we could not identify a pattern as to which genes tend to be correctly predicted by which method.

It is known that the density of genes as well as some of their features, e.g. intron length, depend on the GC contents of the DNA sequence [DMG95, Con01]. To test whether the performance of the methods depends on the GC contents of the input DNA sequences, we subdivided the test set into the following four subsets according to the GC contents intervals defined in [Ber89]. As the GC contents of the two DNA sequences of each pair are well correlated, the DNA sequences were sorted by GC contents in pairs. The four intervals are $gc1 = [0, 0.43)$, comprising four sequence pairs, $gc2 = [0.43, 0.51)$, comprising 22 sequence pairs, $gc3 = [0.51, 0.57)$, comprising 26 sequence pairs, and $gc4 = [0.57, 1]$, comprising 28 sequence pairs. Considering the DOUBLESCAN results without the post-processing step, the sensitivity and specificity for start codons, stop codons, exons and genes show no dependency on the GC contents of the DNA sequences and are the same within statistical errors. The same independence of GC contents was found for GENSCAN. However, in GENSCAN this independence is explicitly established by choosing the model's parameters according to the GC contents of the input DNA sequence, whereas DOUBLESCAN's performance is independent of the GC contents without using GC dependent parameters.

3.3 Prediction of conserved subsequences

DOUBLESCAN without the post-processing step retrieves 69 % of the intergenic subsequences, 48 % of the intron subsequences and 99 % of the exon subsequences as conserved subsequences. The level of conservation in the intergenic subsequences is higher than one would expect for long intergenic subsequences, but can be explained by the fact that the intergenic subsequences of the test set are close to the translation or transcription start and end of the genes where a higher density of conserved subsequences is expected [JBD99].

3.4 Validation of the Stepping Stone algorithm

The Stepping Stone algorithm has been developed in order to accelerate the prediction process as both its time and memory requirement scale essentially linearly with the length of the input sequence. Since it is not guaranteed to find an optimal state path, we compared both the state paths and annotations retrieved by DOUBLESCAN using the Stepping Stone algorithm

to those retrieved by DOUBLESCAN using the Hirschberg algorithm on the test set. For these purposes we consider the DOUBLESCAN results without post-processing as they correspond to the state paths which are to be compared. For 81 % of the DNA sequence pairs, the Stepping Stone algorithm finds the optimal state path (this state path need not be the same as the optimal state path retrieved by the Hirschberg algorithm as there are generally several optimally scoring state paths). Comparing the predicted annotations, 97 % of the predicted genes are the same for both algorithms. The agreement for start codons is 100 % and 98 % for stop codons. At nucleotide level, the agreement for exons is 100 % and 99.8 %, respectively, i.e. close to perfect.

Compared to the annotation, the performance of the Hirschberg algorithm is the same as that of the Stepping Stone algorithm except for a 1 % improvement of the exon sensitivity at exon level and the corresponding 1 % decrease of the rate of overlapping exons.

The average length of the sequences in the test set is around 3300 base pairs and there is on average a BLASTN match every 380 base pairs. If we constrain the Stepping Stone algorithm and Hirschberg algorithm to use the same maximum amount of memory, the prediction process using the Stepping Stone algorithm is on average four times faster than using the Hirschberg algorithm. To give an example, the analysis of one pair of DNA sequences of 9604 base pairs and 10373 base pairs length, respectively, took about 126340 CPU seconds and about 400 MB memory on an Alpha processor with the Hirschberg algorithm, while the analysis with the Stepping Stone algorithm took about 13313 CPU seconds using the same amount of memory. We have used DOUBLESCAN with the Stepping Stone algorithm on pairs of sequences of more than 10^5 base pairs length. As the maximum memory to be used can be set by the user, the memory requirement can be traded for the time requirement and vice versa.

Assuming that the density of BLASTN matches is independent of the sequence length, the gain in time using the Stepping Stone algorithm increases with the length of the DNA sequences to be analysed.

3.5 Summary and discussion

The analysis of a test set of 80 pairs of orthologous mouse and human DNA sequences shows that DOUBLESCAN performs well at gene level and significantly outperforms GENSCAN, the reference non-comparative *ab initio* method. DOUBLESCAN's performance at nucleotide level is high, its sensitivity being 2 % lower and its specificity being 5 % higher than GENSCAN's.

At feature level, DOUBLESCAN'S sensitivity for start and stop codons is slightly higher than GENSCAN's, but its specificity is 11 % and 13 %, respectively, lower specificity than GENSCAN'S. Besides the extra states that help GENSCAN recognise the region 5' of the translation start, it also has an inherent bias towards starting and finishing the state path in intergenic regions and is thus biased towards detecting complete genes comprising start and stop codons. As our test set is entirely composed of DNA sequences which each contain one complete gene, we expect this to help GENSCAN. At exon level, DOUBLESCAN'S sensitivity and specificity are 4 % and 3 %, respectively, lower than GENSCAN's. At gene level, DOUBLESCAN outperforms GENSCAN's sensitivity by 10 % and its specificity by 4 %. One gene which is predicted by DOUBLESCAN and which overlaps the annotated gene is removed in the post-processing step which corresponds to a 1 % rate of missing genes. DOUBLESCAN and GENSCAN agree in more than half of their correctly predicted genes. 72 % of all annotated genes are correctly predicted by one or both of the two methods. DOUBLESCAN and GENSCAN thus complement each other. However, we could not find an obvious pattern that would allow us to predict which genes are correctly identified by which method.

It is interesting that the performance of DOUBLESCAN relative to GENSCAN increases progressively when going from fine scale (nucleotide level) to large scale (gene structure). It appears that long range constraints such as the exon-intron structure of genes can be captured well in the comparative model, even though the detailed modelling is simplified compared to GENSCAN.

The performance of DOUBLESCAN and GENSCAN as reported here for a test set of 80 pairs of orthologous mouse and human DNA sequences each comprising one single complete gene does not permit to conclude that the performance on other test sets, especially long DNA sequences comprising multiple genes, will be the same, see for example [GAA+00a] and [WGM00]. To investigate the performance of DOUBLESCAN on multi gene sequences, pairs of long homologous DNA sequences are needed in which the similarities between the two sequences appear in collinearity. This requirement implies that long semi-articifial DNA sequences comprising several single-gene sequences separated by randomly generated intergenic regions (see for example [GAA+00a]) are not likely to constitute an adequate test for comparative gene prediction methods such as DOUBLESCAN as the level and the patterns of conservation between two homologous intergenic subsequences will not necessarily be similar to those between two randomly generated intergenic subsequences.

Comparing the predictions of the Stepping Stone algorithm and the Hirschberg algorithm, for 81% of the sequence pairs is the state path returned by the Stepping Stone algorithm optimal and 97% of the predicted genes axe identical for the two algorithms. The performance of the Hirschberg algorithm is almost the same as that of the Stepping Stone algorithm, while the gain in time using the Stepping Stone algorithm is significant. This is especially important for the analysis of large genomic sequences for which the Stepping Stone algorithm provides a very efficient practical solution.