Chapter 4

Prediction of mouse and human genes with PROJECTOR

4.1 Introduction and motivation

In Chapter 3 we have shown that DOUBLESCAN can predict related genes given two unannotated DNA sequences as the only input information.

Sometimes, we know more about the pair of input DNA sequences than just their sequence of A, C, G and T letters. One typical example is that we know the genes in one of the two DNA sequences, but not in the other homologous sequence. We then want to find the genes in this sequence given the known genes in the other sequence, i.e. we want to project the annotation of one DNA sequence onto the other DNA sequence whose annotation is not known. To name another example, we may have a set of confirmed introns in both sequences and may want to predict genes in the two sequences under the hypothesis that these introns are true.

In our test set [Pac99], see Section A.2 in Appendix A, the orthologous mouse and human genes are very similar not only at protein level, i.e. comparing their sequences of amino-acids, but also at DNA level. In 97 % of the gene pain, the sequences of amino-acids are encoded on the same number of exons. In 42 % of the gene pain, the sequences of amino-acids are partitioned in the same way into pairs of exons of the same length. For 55 % of the gene pairs, the number of exons is the same, but their lengths are slightly different. The exon-intron structure of related genes is thus very similar concerning the number of exons and their lengths. If we therefore know the gene structure of one input DNA sequence, the related gene in the other input DNA sequence is likely to have the same or a similar number of exons, and

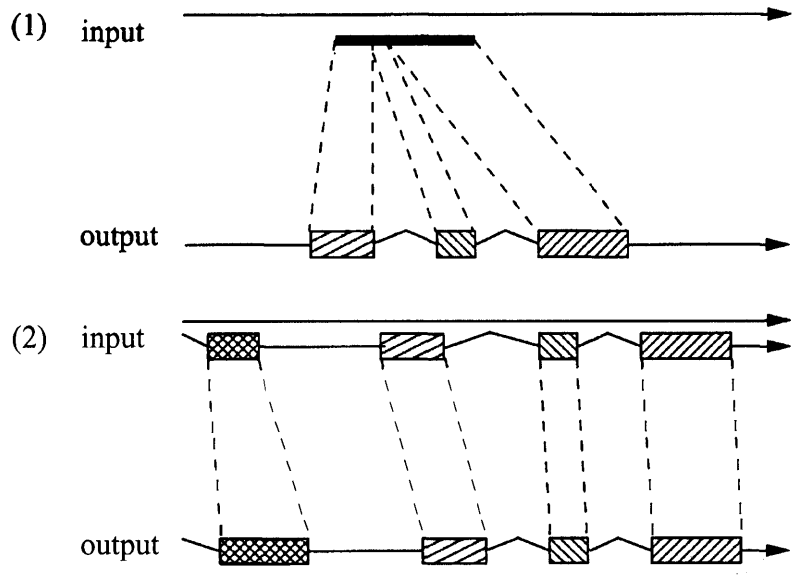

Figure **4.1:** Different types of homology based gene prediction methods: (1) gene prediction based on protein homology (e.g. GENEWISE), (2) gene prediction based on gene homology (e.g. PROJECTOR). Refer to the text for a detailed description of the methods.

the exons will be of the same or a similar length, i.e. the gene structures of the two genes should be very similar. If we knew only the amino-acid sequence of one gene and we would want to find the corresponding related gene in another DNA sequence, we could use programs such as GENEWISE [BD97, BD00] or PROCRUSTES [GMP96], but these methods would a **priori** not know where and if introns are inserted as they lack information on the gene structure. Implementing constraints into the pair HMM of DOUBLESCAN (see Chapter 2), we construct PROJECTOR which can make use of the extra information on the gene structure of one gene to find the gene structure of a related gene. This approach should enable PROJECTOR to find more distantly related genes than is possible with protein based methods [GAA+00b, YLB01].

4.1.1 Implementation

The parameters of the pair HMM according to which the optimal state path is defined, are its transition and emission probabilities, see Section 2.3. By default, they have values which are independent of the positions in the two DNA sequences at which they are used. Some of the transitions within the pair HMM are special, i.e. their values depend on the positions in

the two DNA sequences at which they are used, for example the transitions to the splice site states or to the ***START START*** state, see Section **2.3** and Section **6.2** for a more extensive description. We *can* not only use special transition probabilities, but also special emission probabilities. A state whose emission probabilities are special, reads a score from each of the two sequences depending on the position within that sequence, transforms these scores into posterior probabilities and modifies the nominal value of the emission probability accordingly, *see* Section **6.3** for a detailed description. Special transition and emission probabilities provide the technical concept with which constraints defined by prior knowledge or hypotheses *can* be implemented into the gene prediction.

As an example of how prior knowledge about the input **DNA** sequences can be used within the pair HMM to predict genes, we show how mouse genes *can* be used to predict human genes and vice versa. For this, every state in the pair HMM underlying **PROJECTOR** is defined to have special emission probabilities. Each state can only read letters of the two sequences if the labels of the state match the labels of the sequence whose annotation is used as constraint. To give an example in the case where the annotation of the mouse sequence is used as a constrained to find that of the human sequence: the match exon state has only a non-zero emission probability for reading a pair of codons if the triplet of letters read from the mouse sequences is a codon in the correct phase. The value of the special emission probability thus depends on the letters and the annotation of the mouse DNA sequence at this position, but only the letters of the human DNA sequence. As we know the annotation of one of the two sequences, but not how the two sequences should be aligned, the pair HMM is free to use both match and emit states for finding the optimally scoring state path.

4.2 Results

We have used the mouse human test set of 80 sequence pairs described in Section **A.2** in Appendix A, but discarded three sequence pairs as their annotation cannot be found with the pair HMM. In two pairs, one sequence starts immediately with the start codon whereas the start codon of the other sequence is preceded by a intergenic subsequence. In the third pair, the initial exons consist of the start codon only which cannot be modelled by the states and transitions of the pair HMM. The thus reduced test set of 77 sequence pairs is analysed twice: once, using the human genes to find mouse genes and once using the mouse genes to find the human genes. The pair HMM of Chapter **2** is used, i.e. including the states for UTR-

	Mouse annotation fixed		Human annotation fixed	
	mouse	human	mouse	human
Gene				
Sensitivity	1	0.90	0.90	1
Specificity	1	0.90	0.90	1
Genes overlapping	0	0.10	0.10	0
Genes missing	0	0	0	0
Genes wrong	0	0	0	0
Start Codon				
Sensitivity	1	0.99	0.99	1
Specificity	1	0.99	0.99	1
Stop Codon				
Sensitivity	1	0.96	0.96	1
Specificity	1	0.96	0.96	1
Exon				
Feature Level				
Sensitivity	1	0.97	0.97	1
Specificity	1	0.96	0.97	1
Exons overlapping	0	0.02	0.03	0
Exons missing	0	0.003	0.01	0
Exons wrong	0	0.02	0.01	0
Nucleotide Level				
Sensitivity	1	0.998	0.993	1
Specificity	1	0.995	0.999	1

Table 4.1: Performance figures for PROJECTOR on the mouse human test set. The predictions were generated using the Stepping Stone algorithm. See Table 3.1 for the definitions of rows.

splicing. The results are generated using PROJECTOR with the Stepping Stone algorithm and are shown in Table 4.1. Note that the predicted genes were not post-processed.

The first thing to note is that the performance for predicting entire genes is very high with a sensitivity and specificity of 90 %. The second thing to note is that the performance is symmetric with respect to the two sequences, i.e. it is as difficult to find a human gene given a related mouse gene as it is to find a mouse gene given a related human gene. The ability to detect start codons is almost perfect with a sensitivity and specificity of 99 %, whereas the performance for stop codons is slightly lower with a sensitivity and specificity of 96 %. The sensitivity and specificity for detecting whole exons is about 97 %. At nucleotide level, the

performance for exons is almost perfect.

If we investigate the sixteen genes which were not correctly predicted in detail, we find that fourteen of them are found in pairs, i.e. the mouse gene could not be correctly predicted using the human gene as constraint and vice versa. Four incorrectly predicted genes are found in the two gene pairs for which the number of exons is not the same in the mouse and human gene. These two pairs correspond to the 3 % of the gene pairs in the test set whose genes are related by events of exon-fusion or exon-splitting. In both *cases,* PROJECTOR predicts the wrong number of exons, but not necessarily the same number of exons as in the annotated sequence see Figure 4.2 and Figure 4.3. PROJECTOR'S difficulty in correctly predicting genes which are related by events of exon-fusion or exon-splitting is not surprising as its parameters could not be reliably trained on the single pair of genes of this type within the training set, *see* Section A.1 in Appendix A. Another source of error for six of the sixteen incorrectly predicted genes is the incorrect prediction of a single splice site in an otherwise correctly predicted gene. These incorrectly predicted splice sites are close to the correct ones and introduce no phase shift into the exons, a typical example is shown in Figure 4.4. They may thus correspond to true alternative splice sites. This supposition is fortified by the fact that the incorrectly predicted splice sites are generally not due to PROJECTOR trying to approximate the length of the predicted exon to that of the annotated exon in the other sequence. Four out of the sixteen incorrectly predicted genes are due to a incorrect prediction of the stop codon as shown in Figure 4.5. In one of the sixteen incorrectly predicted genes is a wrong mini exon of *6* base pairs inserted into an otherwise correctly predicted gene, the corresponding pair of genes is shown in Figure 4.6. Two other incorrectly predicted genes are due to incorrectly predicted start codons, see Figure 4.7. In both cases is 'the length of the predicted exon shifted towards the length of the annotated exon in the other sequence without introducing a phase shift. This may be due to a mis-annotation of the start codon in one or other of the sequences, not a failure by PROJECTOR.

4.3 Summary and discussion

PROJECTOR *can* be successfully used to predict genes which are related to known genes and its sensitivity and specificity at gene level is 90 %. Start and stop codons as well as whole exons are predicted with a high reliability as sensitivity and specificity are higher than 95 %. About a third of the incorrectly predicted genes are due to a single splice site being predicted

in close vicinity to the annotated splice site which does not introduce a phase shift into the exons. These cases may correspond to alternative splicing. PROJECTOR's performance could be further improved by training on an enlarged set of pairs of genes which are related by events of exon-splitting or exon-fusion as PROJECTOR so far has difficulty dealing with these cases.

```
Mm.U13921.MK13.1        1-4678  (4678)  forward

 Mm.U13921.MK13. : |AGA-->--480-->--ACC|ATG|AGC-->--468-->--AAG|GTG-->--968-->--TAG|ATT-->--83-->--
 annotation      : |----->--480-->-----|SSS|000-->--468-->--000|----->--968-->-----|000-->--83-->--
 prediction      : |----->--480-->-----|SSS|000-->--468-->--000|----->--968-->-----|000-->--83-->--

 Mm.U13921.MK13. : CAA|GTG-->--194-->--CAG|GTA-->--157-->--GAG|GTG-->--177-->--CAG|GAG-->--162-->--
 annotation      : 000|----->--194-->-----|222-->--157-->--222|----->--177-->-----|000-->--162-->--
 prediction      : 000|----->--194-->-----|222-->--157-->--222|----->--177-->-----|000-->--162-->--

 Mm.U13921.MK13. : AAG|GTA-->--111-->--TAG|AGT-->--126-->--ATG|GTA-->--90-->--CAG|AAA-->--221-->--T
 annotation      : 000|----->--111-->-----|000-->--126-->--000|----->--90-->-----|000-->--221-->--0
 prediction      : 000|----->--111-->-----|000-->--126-->--000|----->--90-->-----|000-->--221-->--0

 Mm.U13921.MK13. : AA|GTA-->--726-->--CAG|GAT-->--23-->--GAG|GTA-->--309-->--CAG|TCTCAG|GAAA|TAA|CA
 annotation      : 00|----->--726-->-----|222-->--23-->--222|----->--309-->-----|------|1111|111|11
 prediction      : 00|----->--726-->-----|----->--23-->-----|----->--309-->-----|222222|2222|SSS|--

 Mm.U13921.MK13. : C-->--61-->--TAT|TAA|CTC-->--303-->--GGC|
 annotation      : 1-->--61-->--111|SSS|----->--303-->-----|
 prediction      : --->--61-->-----|---|----->--303-->-----|

.....................................................................................................

Hs.AF049259.2   1-5575  (5575)  forward

 Hs.AF049259.2   : |GGA-->--511-->--ACC|ATG|AGC-->--492-->--AAG|GTG-->--1324-->--TAG|ATC-->--83-->-
 annotation      : |----->--511-->-----|SSS|000-->--492-->--000|----->--1324-->-----|000-->--83-->-
 prediction      : |----->--511-->-----|SSS|000-->--492-->--000|----->--1324-->-----|000-->--83-->-

 Hs.AF049259.2   : -CAA|GTG-->--199-->--TAG|GTA-->--157-->--GAG|GTG-->--190-->--CAG|GAG-->--162-->-
 annotation      : -000|----->--199-->-----|222-->--157-->--222|----->--190-->-----|000-->--162-->-
 prediction      : -000|----->--199-->-----|222-->--157-->--222|----->--190-->-----|000-->--162-->-

 Hs.AF049259.2   : -AAG|GTA-->--124-->--CAG|AGT-->--126-->--ATG|GTA-->--92-->--CAG|AAA-->--221-->--
 annotation      : -000|----->--124-->-----|000-->--126-->--000|----->--92-->-----|000-->--221-->--
 prediction      : -000|----->--124-->-----|000-->--126-->--000|----->--92-->-----|000-->--221-->--

 Hs.AF049259.2   : CAA|GTA-->--627-->--CAG|GAT-->--26-->--CAG|GTA-->--356-->--CAG|GAA-->--16-->--CC
 annotation      : 000|----->--627-->-----|----->--26-->-----|----->--356-->-----|222-->--16-->--22
 prediction      : 000|----->--627-->-----|222-->--26-->--222|----->--356-->-----|111-->--16-->--11

 Hs.AF049259.2   : G|TAG|CAC-->--85-->--CCT|TAA|ATC-->--775-->--CCA|
 annotation      : 2|SSS|----->--85-->-----|---|----->--775-->-----|
 prediction      : 1|111|111-->--85-->--111|SSS|----->--775-->-----|

-----------------------------------------------------------------------------------------------------
```

Figure **4.2**: One of the two pairs of mouse and human genes that are related by exon-fusion or exon-splitting. The gene of the mouse sequence, **Mm.U13921.MK13.1**, has eight exons whereas the gene of the human sequence, **Hs.AF049259.2,** has seven. The letters of the **DNA** sequence of the forward strand are shown in the upper row, the annotation in the middle row and the prediction generated by **PROJECTOR** in the lower row. Start and stop codons are denoted by **SSS,** letters within exons are denoted according to the exon's phase by 0, 1 or **2** and letters within intron or intergenic regions by **-**. The arrows, **-->--** or **--<--**, indicate the orientation of the **DNA.** The numbers give the length of each segment between two separators (**|**) in base pairs.

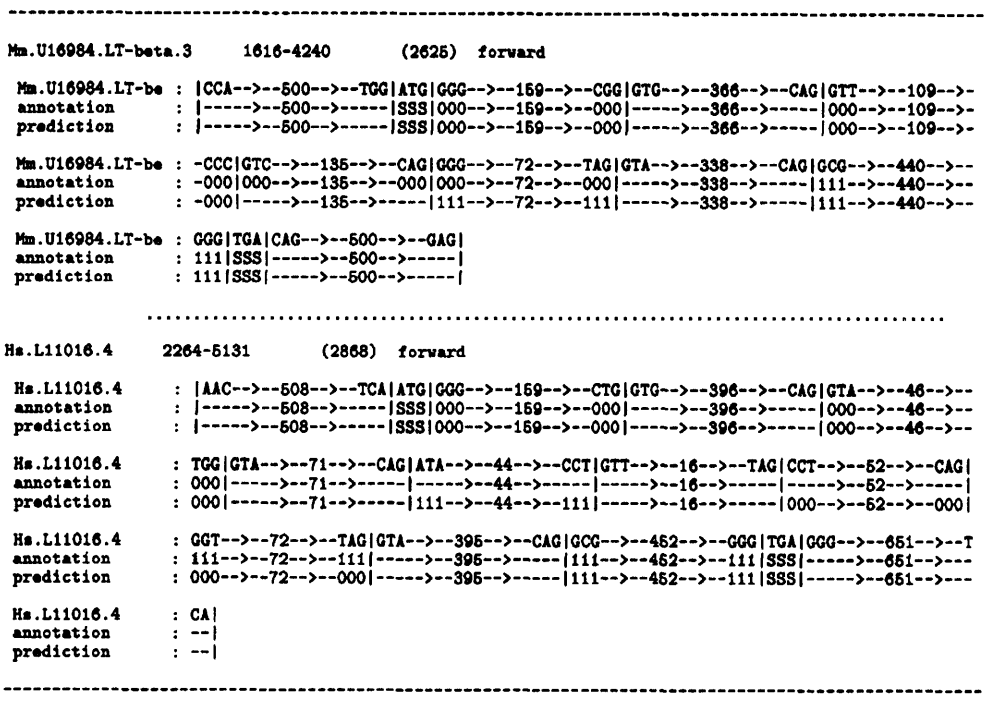

Figure 4.3: The second of the two pairs of mouse and human genes that are related by exon-fusion or exon-splitting. The gene of the mouse sequence, Mm.U16984.LT-beta.3, has three exons whereas the gene of the human sequence, Hs.L11016.4, has four. See Figure 4.2 for an explanation of the notation.

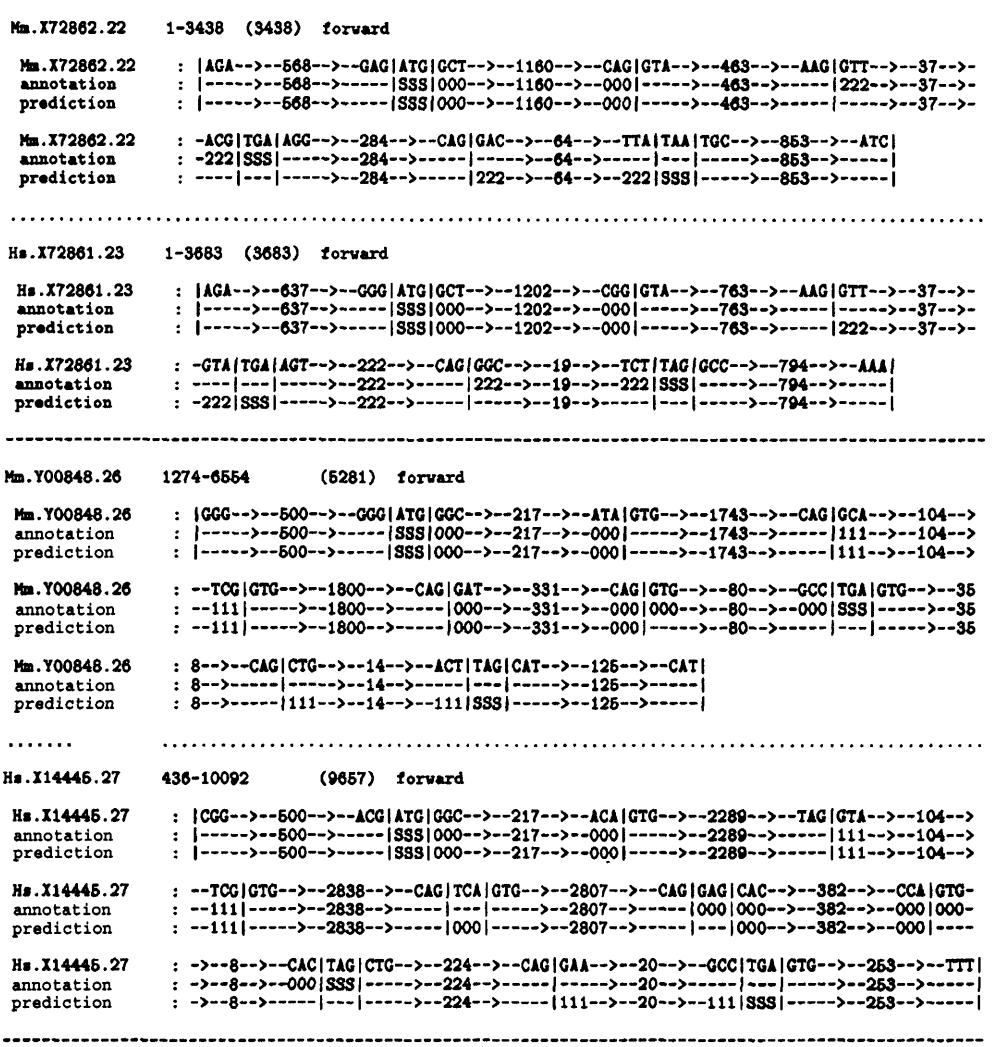

Figure 4.5: The two pairs of genes whose stop codons were incorrectly predicted. The names of the mouse sequences start with Mm, those of the human sequences with Hs. *See* Figure 4.2 for an explanation of the notation.

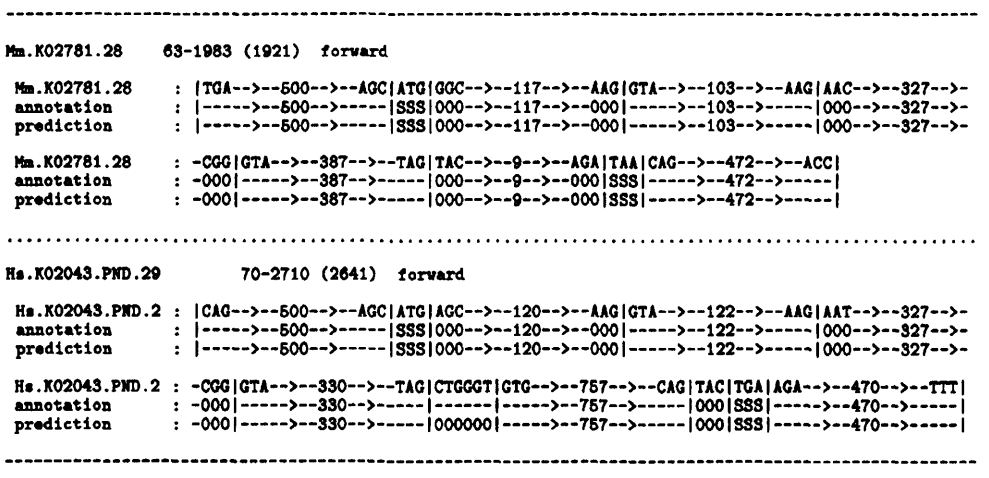

Figure 4.6: This prediction of the human gene, see sequence Hs.K02043.PND.29, contains a wrong mini exon of six base pairs length. The corresponding mouse gene, Mm.K02781.28, is also shown for comparison. See Figure 4.2 for an explanation of the notation.