Chapter 5

Prediction of *C. elegans* and C. *briggsae* genes using DOUBLESCAN and PROJECTOR

5.1 Introduction and motivation

The genome of the nematode Caenorhabditis elegans was the first multi-cellular organism to be sequenced in 1998 [eSC98]. This model organism has been studied in great detail: we know its developmental lineage to the cellular level and even the entire wiring diagram of its nerve cells. Databases are being maintained which aim at integrating all available information from experimental and theoretical studies into a single coherent picture of the organism [MBD97, SSD+01, Wor]. The sequencing of a related nematode, Caenorhabditis briggsae, is now near to completion. C. elegans and C. briggsae are estimated to have diverged from a common ancestor around 25–100 million years ago, see for example [KAA+93, BFV+97, VPS98] (these estimates vary greatly as they rely on a number of assumptions about mutation rates which cannot be verified as fossil records are not available [ABK96]). The comparative large scale analysis of these two genomes will deepen and maybe also revise our current understanding of the C. elegans genome and at the same time provide the C. briggsae genome with a first global annotation. These analyses will not only aim at detecting the protein coding genes of the two genomes, but will also investigate short conserved regions which may have regulatory functions as well as the large scale structure of the two genomes.

Our main motivation for studying C. elegans and C. briggsae DNA sequences is to test if the pair HMM underlying DOUBLESCAN and PROJECTOR can be easily adapted to successfully predict genes in other pairs of related genomes. The non-comparative ab **initio** gene prediction method GENSCAN which is the reference method for the ab **initio** prediction of human genes, **was** reported to have a 'rather poor performance for C. *elegans* genomic sequences' [Bur97, pp. 107] which was attributed to the difficulty of its gene model in dealing with nematode specific features such **as** trans-splicing. In designing the pair HMM underlying DOUBLESCAN and PROJECTOR, the main idea was to keep the gene model **as** general **as** possible so that it can be **used** on any pair of related eukaryotic genomes and to introduce the specialisation to a certain pair of genomes only through the parameters of the model which should be either robust or easily adaptable to new data.

5.2 Training of the pair HMM's parameters

The architecture of the pair HMM underlying DOUBLESCAN and PROJECTOR **as** described in Chapter 2 is suitable for the prediction of genes in any pair of related eukaryotic organisms. The specialisation for a certain pair of genomes is only introduced by setting its parameters accordingly.

The pair HMM was adapted to analyze DNA sequences of C. elegans and C. briggsae instead of mouse and human by implementing the following changes:

- Both the parametrisation of the transition probabilities and the values of the parameters (see Table B.1 and Table B.2 in Appendix B) are exactly the same **as** for the mouse human analysis. Only the prior for the transition from an intron state to a **3'** splice site was increased from 1/1000 to 1/100 **as** introns within C. elegans and C. briggsae are on average an order of magnitude shorter than in mouse and human introns (compare the values **for** Prior_AG in Table **B.3** in Appendix **B** and Table D.1 in Appendix D).
- The emission probabilities of the pair HMM were automatically derived from a training set [1] of known C. elegans C. briggsae gene pairs (see Section C.1 in Appendix C) in the same way **as** they were derived from a training set of known mouse human gene pairs, see Section 2.3 and Section 2.3 **for** a detailed description.

[1] **I thank Avril Coghlan, Trinity College, Dublin, for the preparation of the training set.**

- The splice sites scores and start codon scores described in Section **2.3** are generated by GENEFINDER [eSC98], a program which was trained on C. *elegans* genes, rather than by STRATASPLICE [LD01] which is used for the mouse and human analysis. GENEFINDER is used with default cutoff values (**−2** for **3'** splice sites and 0 for 5' splice sites and start codons) ***so*** that only splice sites and start codons which score above these cutoff values are taken into account by DOUBLESCAN and PROJECTOR, whereas for the mouse human analysis every potential translation start site and splice site is considered. In addition to the consensus splice sites GT at the 5' splice site and **AG** at the 3' splice site which are the only splice sites considered for the mouse human analysis, also GC is enabled for 5' splice sites **as** this type of **5'** splice site occurs with a frequency of about 10 % in C. *elegans* introns. This frequency is similar for mouse and human introns, but we chose to model non-consensus splice sites only for nematodes **as** we expect the performance for nematodes to be ***so*** good that this effect will be relevant.

The above changes are the only changes made in order to transform the pair HMM underlying DOUBLESCAN and PROJECTOR from a mouse human into a C. *elegans* C. *briggsae* gene prediction program. In particular, its transition probabilities were not tuned by hand to further optimise the performance.

5.3 Results

DOUBLESCAN and PROJECTOR are run with the Stepping Stone algorithm on the two test sets of C. *elegans* and *C. briggsae* DNA sequences (see Section **C.2** and Section **C.3** in Appendix **C)**. **As** for the analyses of mouse and human DNA sequences described in Chapter **3** and Chapter **4,** the DNA sequences are not masked for repeats **or** anything else. The gene prediction is done three times. Once, using DOUBLESCAN to predict genes simultaneously in C. *elegans* and *C. briggsae* in an *ab initio* way, once keeping the annotation of the C. *elegans* sequences fixed to find C. *briggsae* genes using PROJECTOR and once keeping the annotation of the C. *briggsae* sequences fixed to find C. *elegans* genes. The predicted genes generated **by** DOUBLESCAN are compared to the annotated genes. The set of predicted genes is not post-processed. The results of the comparison are shown in Table 5.1. The columns labelled 'PROJECTOR' contain the performance on the joint set of C. *elegans* genes which are predicted by keeping the C. *briggsae* genes fixed and the C. *briggsae* genes which are predicted by keeping the C. *elegans* genes **fixed.**

	Test set 1		Test set 2	
	DOUBLESCAN	PROJECTOR	DOUBLESCAN	PROJECTOR
Gene				
Sensitivity	0.80	0.95	0.74	0.90
Specificity	0.71	0.95	0.62	0.90
Genes overlapping	0.23	0.05	0.28	0.10
Genes missing	0	0	0.01	0
Genes wrong	0.06	0	0.10	0
Start Codon				
Sensitivity	0.96	0.99	0.96	0.99
Specificity	0.87	0.99	0.81	0.99
Stop Codon				
Sensitivity	0.96	0.997	0.93	0.99
Specificity	0.89	0.997	0.82	0.99
Exon				
Feature Level				
Sensitivity	0.93	0.99	0.91	0.97
Specificity	0.90	0.98	0.89	0.97
Exons overlapping	0.06	0.01	0.07	0.02
Exons missing	0.004	0.003	0.02	0.003
Exons wrong	0.04	0.01	0.04	0.01
Nucleotide Level				
Sensitivity	0.996	0.997	0.98	0.995
Specificity	0.991	0.998	0.99	0.998

Table 5.1: Performance figures for DOUBLESCAN and PROJECTOR on the two C. *elegans* and C. *briggsae* test sets. The predictions by DOUBLESCAN and PROJECTOR were generated using the Stepping Stone algorithm. The table does not include the performance on the C. *elegans* and C. *briggsae* sequences separately as they are very similar. See Table 3.1 for the definitions of rows.

	Test set 1				Test set 2			
DOUBLESCAN								
incorrectly predicted genes	139 (23 %)				268 (28 %)			
source of error	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
split genes	36		26	26	88		33	27
incorrect or missing start codons	31	19	22	22	37	37	14	11
incorrect or missing stop codons	30		22	22	69		26	21
incorrectly predicted splice sites	24	20	17	17	70	64	26	22
wrong exons	14	11	10	10	39	36	15	12
missing introns	3	2	2	2	10	10	4	3
missing exons	2	2	1	1	8	8	3	3
inserted introns	0	0	0	0	4	4	1	1
sum	140			100	325			100
PROJECTOR								
incorrectly predicted genes	36 (5 %)				112 (10 %)			
source of error	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
incorrectly predicted splice sites	12	9	33	31	63	51	56	51
wrong exons	12	11	33	31	28	26	25	23
incorrect start codons	8	5	22	20	5	4	4	4
missing introns	3	3	8	8	13	13	12	10
missing exons	2	2	6	5	2	2	2	2
incorrectly predicted stop codons	2		6	5	10		9	8
inserted introns	0	0	0	0	3	2	3	2
sum	39			100	124			100

Table 5.2: Error analysis for the genes of the two C. *elegans* and C. *briggsae* test sets which are incorrectly predicted by DOUBLESCAN or PROJECTOR. Column (1) gives the number of incorrectly predicted genes with this type of error, column (2) gives the number of incorrectly predicted genes where this type of error does not lead to a phase shift, column (3) gives the percentage of incorrectly predicted genes with this error and column (4)the percentage of this error within all errors. To give an example: PROJECTOR predicts 36 genes incorrectly which corresponds to 5 % of the annotated genes in test set 1. 12 of the 36 incorrectly predicted genes have incorrectly predicted splice sites, but this leads in 9 out of 12 genes to no phase shift. 33 % of incorrectly predicted genes have an incorrectly predicted splice site and incorrectly predicted splice sites correspond to 31 % of the errors made. Note that the sum of numbers in column (1)need not be equal to the number of incorrectly predicted genes and the sum of numbers in column (3) is not necessarily 100 % as some genes are affected by more than one type of error.

The first thing to note is that the performance both of the ab initio gene prediction and the homology based prediction is very good. This is very promising, especially given the fact that the switch from the mouse and human to the C. eiegans and C. briggsae pair HMM which underlies **DOUBLESCAN** and **PROJECTOR** consists only of a few steps. The second thing to note is that the performance on test set **1** is significantly better than that on test set **2** both in terms of sensitivity and specificity. Note that the table does not include the performance on the C. elegans and C. briggsae sequences separately **as** they are very similar. **DOUBLESCAN** is not biased towards preferentially predicting C. eiegans or C. briggsae genes correctly.
As the two test sets have been generated in different ways, they are discussed separately.

5.3.1 Performance on test set 1

Performance of the ab initio **gene prediction with DOUBLESCAN** Though the sensitivity of the prediction is generally high with **80** % at gene level, the specificity on gene and feature level is generally significantly lower, but the sensitivity and specificity values converge when going from gene level to nucleotide level. **DOUBLESCAN** detects start and stop codons with **95 %** specificity and its sensitivity and specificity for whole exons are above **90** %.
The set of **139** incorrectly predicted genes which overlap an annotated gene ***can*** be subdivided into subsets according to the error that was made, see Table **5.2** for an overview. There are three main errors.
The first type of error in **36** out of the **139** genes consists of splitting the gene into two (or three in six cases) genes which overlap the annotated gene. The overlap between the predicted genes and the annotated gene is generally very large and the split typically involves only two incorrectly predicted splice sites, see Figure **5.1** for an example.
The next common type of error present in **31** out of the **139** incorrectly predicted genes is a start codon which is incorrectly predicted or missing in the predicted gene. An incorrectly predicted start codon **(15** out of **31)** is typically close to the annotated one and does not lead to a phase shift **(13** out of **15).** A typical example is shown in Figure **5.2.** If the start codon is missing from the prediction **(16** out of **31**), there is usually a splice site predicted in close vicinity to the annotated start codon, but this splice site can **(10** out of **16)** or cannot **(6** out of **16)** lead to a phase shift. Figure **5.3** shows an example in which the missing start codon does not lead to a frame shift.
Another common type of error shown in **30** out of the **139** incorrectly predicted genes is a stop

number of amino-acids	number of genes
3	8
4	5
5	4
6	6
7	2
8	2
12	1
18	2
27	1
30	1

Table 5.3: Length distribution of the 32 wrong complete genes predicted by DOUBLESCAN on test set 1. All genes are single exon genes.

codon which is incorrectly predicted (2 out of 30) or missing (28 out of 30) in the predicted gene. As for missing start codons, a missing stop codon is usually due to a splice site being predicted close the the annotated stop codon. A typical example can be seen in Figure 5.4.

The rest of the errors found in the remaining 42 of the 139 incorrectly predicted genes are due to incorrectly predicted splice sites (24 cases), extra wrong exons being predicted (14 cases), an intron missing in the predicted gene (3 cases) or an exon missing in the predicted gene (2 cases). Twenty of the 24 genes in which a splice site is incorrectly predicted do not lead to phase shifts and the predicted splice site is close to the annotated one. These cases may thus be real splice sites which are used in alternative splicing. See Figure 5.5 for an example. The 14 genes in which an extra wrong exon has been predicted are mainly (11 out of 14) due to short exons which do not introduce a phase shift as their length is a multiple of three, see Figure 5.6.

The 6 % rate of wrong genes corresponds to 50 genes which do not overlap any annotated gene. They consist of 32 complete and 18 partial genes. The complete genes are typically very short, see Table 5.3, and are all single exon genes. The partial genes consist of a partial intron, an exon and the start or stop codon. 40 % of the wrong genes lie 5' to the annotated gene and 60 % lie 3' to the annotated gene.

Performance of the gene prediction with PROJECTOR The performance of **PROJECTOR** is very high with a sensitivity and specificity of **95** % at gene level. As the pair HMM underlying **DOUBLESCAN** and **PROJECTOR** predicts genes in pairs, the rate of missing and wrong genes for **PROJECTOR** is zero by construction. The low percentage of overlapping genes corresponds to **36** genes.

The two main sources of errors are incorrectly predicted splice sites and wrong intermediate exons of short length. There are **12** genes with incorrectly predicted splice sites which all do not lead to an overall phase shift. In nine out of **12** genes the incorrectly predicted splice sites do not changes the phase and in the other three the two incorrectly predicted splice sites follow each other and have no overall phase shifting effect. The incorrect splice sites are typically close to the annotated ones and may correspond to true splice sites which may be mis-annotated or used in alternative splicing. Twelve of the incorrectly predicted genes are due to the prediction of a wrong intermediate exon of short length. Almost all of them **(11** out of **12)** do not lead to a phase shift and thus correspond at protein level to the insertion of few amino-acids. The next common error present in eight out of the **36** incorrectly predicted genes are incorrectly predicted start codons. In five out of the eight cases there is no phase shift due to the incorrectly predicted start codon. The remaining errors are due to **missing** introns that do not alter the phase of the exons **(3** genes), missing exons **(2** genes) and incorrectly predicted stop codons **(2** genes).

5.3.2 Performance on test set 2

Test set **2** consists of more diverged pairs of genes (see Table **C.3** in Appendix **C)** whose genes have on average more exons and are longer than those of test set **1** (see Table **C.l** in Appendix **C).**

Performance of the ab **initio gene prediction with DOUBLESCAN** Sensitivity and specificity at gene level on test set **2** are generally lower than on test set **1,** the sensitivity of 74 % being **6** % lower and the specificity of **62** % being **9** % lower. Still, two thirds of the genes are perfectly predicted which is very high for an ***ab* initio** method. As **for** test set **1,** the values for sensitivity and specificity converge when going from gene level to nucleotide level performance where they are almost the same. Both sensitivity and specificity for whole exons are around 90 % and the sensitivity for detecting start and stop codons is even higher.

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

CE.C06G1.1.f     16658735-16668767        (10033) forward

 CE.C06G1.1.f      : |CGA-->--5433-->--CTG|ATG|ATT-->--61-->--CAA|GTG-->--58-->--CAG|ATA-->--158-->--
 annotation        : |----->--5433-->-----|SSS|000-->--61-->--000|----->--58-->-----|111-->--158-->--
 prediction        : |----->--5433-->-----|SSS|000-->--61-->--000|----->--58-->-----|111-->--158-->--

 CE.C06G1.1.f      : CAG|GTA-->--383-->--CAG|GTA-->--99-->--ACG|GTA-->--50-->--CAG|GGA-->--92-->--CAA
 annotation        : 111|----->--383-->-----|000-->--99-->--000|----->--50-->-----|000-->--92-->--000
 prediction        : 111|----->--383-->-----|000-->--99-->--000|----->--50-->-----|000-->--92-->--000

 CE.C06G1.1.f      : |GTG-->--165-->--CAG|AAC-->--181-->--CAG|GTA-->--70-->--CAG|GTT-->--115-->--TCG|
 annotation        : |----->--165-->-----|222-->--181-->--222|----->--70-->-----|000-->--115-->--000|
 prediction        : |----->--165-->-----|222-->--181-->--222|----->--70-->-----|000-->--115-->--000|

 CE.C06G1.1.f      : GTA-->--359-->--CAG|GAC-->--72-->--AAG|GTA-->--70-->--CAG|GAA-->--252-->--CCG|GT
 annotation        : ----->--359-->-----|111-->--72-->--111|----->--70-->-----|111-->--252-->--111|--
 prediction        : ----->--359-->-----|111-->--72-->--111|----->--70-->-----|111-->--252-->--111|--

 CE.C06G1.1.f      : A-->--200-->--CAG|CAC-->--128-->--CAG|GTA-->--97-->--CAG|CAA-->--126-->--GAT|GTA
 annotation        : --->--200-->-----|111-->--128-->--111|----->--97-->-----|000-->--126-->--000|---
 prediction        : --->--200-->-----|111-->--128-->--111|----->--97-->-----|000-->--126-->--000|---

 CE.C06G1.1.f      : -->--49-->--CAG|GTT-->--153-->--AAT|GTC-->--13-->--AAG|AT|G|TAG|GTA-->--780-->--
 annotation        : -->--49-->-----|000-->--153-->--000|000-->--13-->--000|00|-|---|----->--780-->--
 prediction        : -->--49-->-----|000-->--153-->--000|----->--13-->-----|00|0|SSS|----->--780-->--

 CE.C06G1.1.f      : CAG|GAC|ATG|CTG-->--96-->--GGT|TAA|AGT-->--48-->--GTG|ATG|AACG|GTG-->--700-->--T
 annotation        : ---|000|000|000-->--96-->--000|SSS|----->--48-->-----|---|----|----->--700-->---
 prediction        : ---|---|SSS|000-->--96-->--000|SSS|----->--48-->-----|SSS|0000|----->--700-->---

 CE.C06G1.1.f      : TT|
 annotation        : --|
 prediction        : --|

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

CB.gf.s146.9.r   35800-40603      (4804)   reverse

 CB.gf.s146.9.r   : |CAA-->--476-->--CAC|CGTT|CAT|CAC-->--56-->--TTT|TTA|ACC-->--96-->--CAC|CAT|GTC|
 annotation       : |-----<--476--<-----|----|---|-----<--56--<-----|SSS|000--<--96--<--000|000|000|
 prediction       : |-----<--476--<-----|0000|SSS|-----<--56--<-----|SSS|000--<--96--<--000|SSS|---|

 CB.gf.s146.9.r   : CTG-->--180-->--TAC|TCA|C|AT|CTT-->--13-->--AAC|ATT-->--153-->--GAC|CTG-->--52--
 annotation       : -----<--180--<-----|---|-|00|000--<--13--<--000|000--<--153--<--000|-----<--52--
 prediction       : -----<--180--<-----|SSS|0|00|-----<--13--<-----|000--<--153--<--000|-----<--52--

 CB.gf.s146.9.r   : >--TAC|ATC-->--126-->--TTG|CTG-->--96-->--TAC|CTG-->--128-->--GAG|CTG-->--50-->-
 annotation       : <-----|000--<--126--<--000|-----<--96--<-----|111--<--128--<--111|-----<--50--<-
 prediction       : <-----|000--<--126--<--000|-----<--96--<-----|111--<--128--<--111|-----<--50--<-

 CB.gf.s146.9.r   : -TAC|CAG-->--252-->--TTC|CTG-->--72-->--TAC|CTT-->--72-->--GGC|CTG-->--292-->--T
 annotation       : ----|111--<--252--<--111|-----<--72--<-----|111--<--72--<--111|-----<--292--<---
 prediction       : ----|111--<--252--<--111|-----<--72--<-----|111--<--72--<--111|-----<--292--<---

 CB.gf.s146.9.r   : AC|CTA-->--115-->--AAC|CTG-->--42-->--CAC|CTG-->--181-->--GTT|CTG-->--192-->--CA
 annotation       : --|000--<--115--<--000|-----<--42--<-----|222--<--181--<--222|-----<--192--<----
 prediction       : --|000--<--115--<--000|-----<--42--<-----|222--<--181--<--222|-----<--192--<----

 CB.gf.s146.9.r   : C|TTG-->--92-->--ACC|CTG-->--43-->--TAC|TGT-->--99-->--AAC|CTG-->--48-->--TAC|TT
 annotation       : -|000--<--92--<--000|-----<--43--<-----|000--<--99--<--000|-----<--48--<-----|11
 prediction       : -|000--<--92--<--000|-----<--43--<-----|000--<--99--<--000|-----<--48--<-----|11

 CB.gf.s146.9.r   : G-->--158-->--TGT|CTG-->--48-->--CAC|TTG-->--61-->--AAT|CAT|CAG-->--1586-->--AAG|
 annotation       : 1--<--158--<--111|-----<--48--<-----|000--<--61--<--000|SSS|-----<--1586--<-----|
 prediction       : 1--<--158--<--111|-----<--48--<-----|000--<--61--<--000|SSS|-----<--1586--<-----|

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

Figure 5.1: Example of a gene pair where the annotated gene is split into two genes predicted by DOUBLESCAN which overlap the annotated gene. The prediction also contains two partial genes which are wrong. The C. elegans sequence, CE.C06G1.1.f, is shown at the top, the corresponding C. *briggsae* sequence, CB.gf.s146.9.r, at the bottom. See Figure 4.2 for an explanation of the notation.

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

CE.C25H3.9.r      5689738-5691493 (1756)  reverse

 CE.C25H3.9.r     : |CAA-->--864-->--CAA|TTA|CTG-->--154-->--TGC|CTG-->--45-->--TAC|CTC-->--140-->--
 annotation       : |-----<--864--<-----|SSS|222--<--154--<--222|-----<--45--<-----|000--<--140--<--
 prediction       : |-----<--864--<-----|SSS|222--<--154--<--222|-----<--45--<-----|000--<--140--<--

 CE.C25H3.9.r     : TCC|CTG-->--49-->--TAC|TTC-->--111-->--ATC|CTT-->--47-->--TAC|CTT-->--132-->--CG
 annotation       : 000|-----<--49--<-----|000--<--111--<--000|-----<--47--<-----|000--<--132--<--00
 prediction       : 000|-----<--49--<-----|000--<--111--<--000|-----<--47--<-----|000--<--132--<--00

 CE.C25H3.9.r     : C|CAT|TTT-->--15-->--CGC|CAT|GTT-->--190-->--AAG|
 annotation       : 0|SSS|-----<--15--<-----|---|-----<--190--<-----|
 prediction       : 0|000|000--<--15--<--000|SSS|-----<--190--<-----|

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

CB.gf.s150.69.r 238584-240007   (1424)  reverse

 CB.gf.s150.69.r : |ATG-->--544-->--CAT|TCA|CTG-->--154-->--TGT|CTA-->--49-->--TAC|CTC-->--140-->--
 annotation      : |-----<--544--<-----|SSS|222--<--154--<--222|-----<--49--<-----|000--<--140--<--
 prediction      : |-----<--544--<-----|SSS|222--<--154--<--222|-----<--49--<-----|000--<--140--<--

 CB.gf.s150.69.r : TCC|CTG-->--50-->--TAC|CTC-->--111-->--GTC|CTG-->--51-->--TAC|CTT-->--132-->--AG
 annotation      : 000|-----<--50--<-----|000--<--111--<--000|-----<--51--<-----|000--<--132--<--00
 prediction      : 000|-----<--50--<-----|000--<--111--<--000|-----<--51--<-----|000--<--132--<--00

 CB.gf.s150.69.r : C|CAT|TTT-->--15-->--CGC|CAT|GGT-->--169-->--GTC|
 annotation      : 0|SSS|-----<--15--<-----|---|-----<--169--<-----|
 prediction      : 0|000|000--<--15--<--000|SSS|-----<--169--<-----|

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

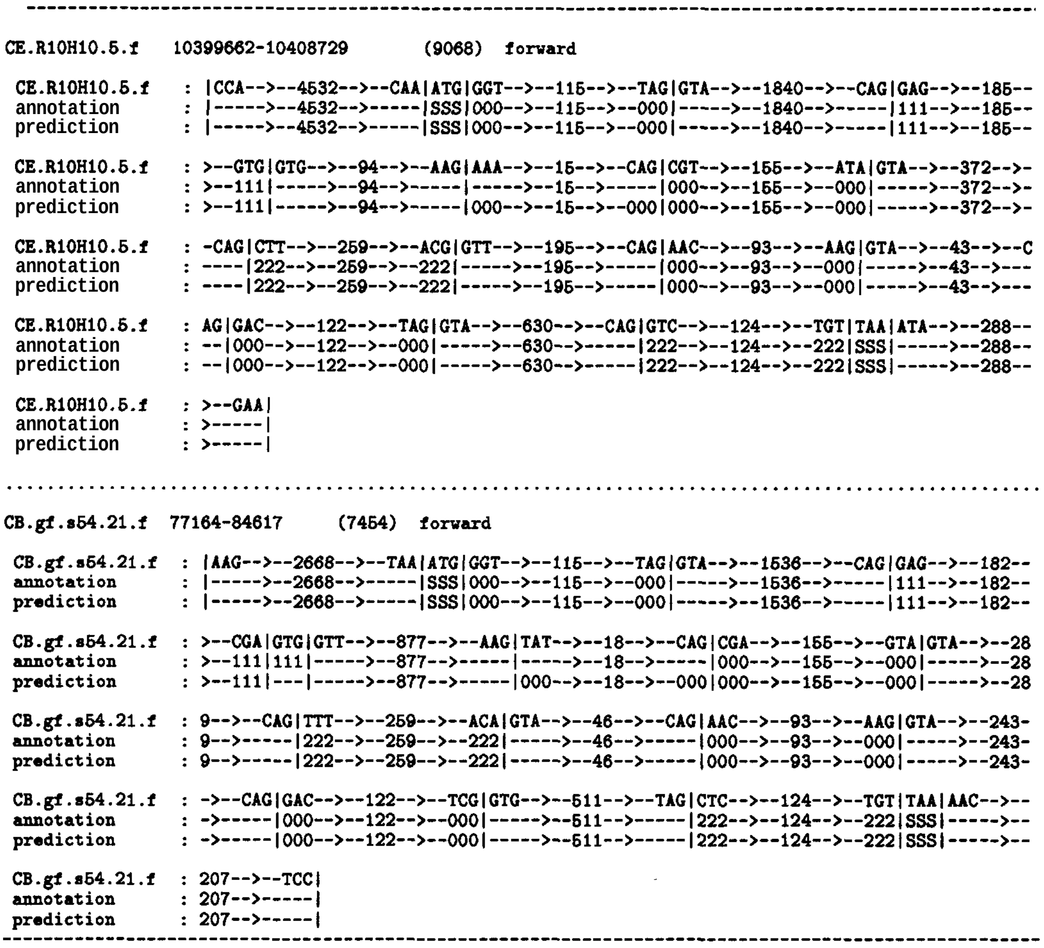
Figure 5.2: Example of a gene pair where the start codon of the gene predicted by **DOUBLESCAN** lies close to the annotated one and involves no phase shift. Note that the genes in this example lie on the reverse strand. The C. elegans sequence, CE.C25H3.9.r, is shown at the top, the corresponding C. briggsae sequence, CB.gf.s150.69.r, at the bottom. See Figure 4.2 for an explanation of the notation.

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

CE.F35G2.2.r     12207941-12209531        (1591)  reverse

 CE.F35G2.2.r     : |TTC-->--532-->--CAT|TTA|GAG-->--339-->--AAA|CTG-->--51-->--TAC|TTT-->--178-->--
 annotation       : |-----<--532--<-----|SSS|000--<--339--<--000|-----<--51--<-----|222--<--178--<--
 prediction       : |-----<--532--<-----|SSS|000--<--339--<--000|-----<--51--<-----|222--<--178--<--

 CE.F35G2.2.r     : GAC|CTG-->--57-->--TAC|GCG-->--77-->--AGC|CAT|CAC|CTG-->--348-->--CAA|
 annotation       : 222|-----<--57--<-----|000--<--77--<--000|SSS|---|-----<--348--<-----|
 prediction       : 222|-----<--57--<-----|000--<--77--<--000|000|000|-----<--348--<-----|

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

CB.gf.s6.24.f    71004-72828      (1825)  forward

 CB.gf.s6.24.f    : |AAA-->--630-->--CAG|ATC|ATG|GCT-->--77-->--TGC|GTA-->--50-->--CAG|GTC-->--178--
 annotation       : |----->--630-->-----|---|SSS|000-->--77-->--000|----->--50-->-----|222-->--178--
 prediction       : |----->--630-->-----|000|000|000-->--77-->--000|----->--50-->-----|222-->--178--

 CB.gf.s6.24.f    : >--AAG|GTA-->--54-->--TAG|TTC-->--339-->--CTC|TAA|ATT-->--488-->--AGT|
 annotation       : >--222|----->--54-->-----|000-->--339-->--000|SSS|----->--488-->-----|
 prediction       : >--222|----->--54-->-----|000-->--339-->--000|SSS|----->--488-->-----|

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

Figure 5.3: Example **of** a gene pair where the start codons are missing in the genes predicted by **Doublescan. A** splice site has been introduced in close vicinity to the annotated start codon which does not lead to a phase shift. The C. *elegans* sequence, CE.F35G2.2.r, with the gene lying on the reverse strand is shown at the top, the corresponding C. briggsae sequence, CB.gf.s6.24.f, at the bottom. See Figure **4.2 for** an explanation **of** the notation.

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

CE.C06B8.8.f     15514309-15515518        (1210)  forward

 CE.C06B8.8.f     : |ACA-->--481-->--GCG|ATG|CCA-->--141-->--GCC|GTA-->--272-->--CAG|GAC-->--65-->--
 annotation       : |----->--481-->-----|SSS|000-->--141-->--000|----->--272-->-----|000-->--65-->--
 prediction       : |----->--481-->-----|SSS|000-->--141-->--000|----->--272-->-----|000-->--65-->--

 CE.C06B8.8.f     : CAA|G|TAG|ATC-->--244-->--GTT|
 annotation       : 000|0|SSS|----->--244-->-----|
 prediction       : 000|-|---|----->--244-->-----|

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

CB.gf.s219.10.f 53654-56301      (2648)  forward

 CB.gf.s219.10.f : |AGT-->--1786-->--ACA|ATG|CCA-->--141-->--GCC|GTG-->--553-->--CAG|GAC-->--65-->-
 annotation      : |----->--1786-->-----|SSS|000-->--141-->--000|----->--553-->-----|000-->--65-->-
 prediction      : |----->--1786-->-----|SSS|000-->--141-->--000|----->--553-->-----|000-->--65-->-

 CB.gf.s219.10.f : -CAA|G|TAG|ATT-->--96-->--AAT|
 annotation      : -000|0|SSS|----->--96-->-----|
 prediction      : -000|-|---|----->--96-->-----|

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

Figure **5.4:** Example **of** a gene pair where the stop codons are missing in the genes predicted by **Doublescan. A** splice site has been introduced in close vicinity to the annotated stop codon. The C. elegans sequence, CE.C06B8.8.f, is shown at the top, the corresponding C. briggsae sequence, CB.gf.s219.10.f, at the bottom. *See* Figure **4.2 for** an explanation **of** the notation.



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

CE.R10H10.5.f   10399662-10408729        (9068)  forward

 CE.R10H10.5.f     : |CCA-->--4532-->--CAA|ATG|GGT-->--115-->--TAG|GTA-->--1840-->--CAG|GAG-->--185--
 annotation        : |----->--4532-->-----|SSS|000-->--115-->--000|----->--1840-->-----|111-->--185--
 prediction        : |----->--4532-->-----|SSS|000-->--115-->--000|----->--1840-->-----|111-->--185--

 CE.R10H10.5.f     : >--GTG|GTG-->--94-->--AAG|AAA-->--15-->--CAG|CGT-->--155-->--ATA|GTA-->--372-->-
 annotation        : >--111|----->--94-->-----|----->--15-->-----|000-->--155-->--000|----->--372-->-
 prediction        : >--111|----->--94-->-----|000-->--15-->--000|000-->--155-->--000|----->--372-->-

 CE.R10H10.5.f     : -CAG|CTT-->--259-->--ACG|GTT-->--195-->--CAG|AAC-->--93-->--AAG|GTA-->--43-->--C
 annotation        : ----|222-->--259-->--222|----->--195-->-----|000-->--93-->--000|----->--43-->---
 prediction        : ----|222-->--259-->--222|----->--195-->-----|000-->--93-->--000|----->--43-->---

 CE.R10H10.5.f     : AG|GAC-->--122-->--TAG|GTA-->--630-->--CAG|GTC-->--124-->--TGT|TAA|ATA-->--288--
 annotation        : --|000-->--122-->--000|----->--630-->-----|222-->--124-->--222|SSS|----->--288--
 prediction        : --|000-->--122-->--000|----->--630-->-----|222-->--124-->--222|SSS|----->--288--

 CE.R10H10.5.f     : >--GAA|
 annotation        : >-----|
 prediction        : >-----|

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

CB.gf.s54.21.f  77164-84617      (7454)  forward

 CB.gf.s54.21.f    : |AAG-->--2668-->--TAA|ATG|GGT-->--115-->--TAG|GTA-->--1536-->--CAG|GAG-->--182--
 annotation        : |----->--2668-->-----|SSS|000-->--115-->--000|----->--1536-->-----|111-->--182--
 prediction        : |----->--2668-->-----|SSS|000-->--115-->--000|----->--1536-->-----|111-->--182--

 CB.gf.s54.21.f    : >--CGA|GTG|GTT-->--877-->--AAG|TAT-->--18-->--CAG|CGA-->--155-->--GTA|GTA-->--28
 annotation        : >--111|111|----->--877-->-----|----->--18-->-----|000-->--155-->--000|----->--28
 prediction        : >--111|---|----->--877-->-----|000-->--18-->--000|000-->--155-->--000|----->--28

 CB.gf.s54.21.f    : 9-->--CAG|TTT-->--259-->--ACA|GTA-->--46-->--CAG|AAC-->--93-->--AAG|GTA-->--243-
 annotation        : 9-->-----|222-->--259-->--222|----->--46-->-----|000-->--93-->--000|----->--243-
 prediction        : 9-->-----|222-->--259-->--222|----->--46-->-----|000-->--93-->--000|----->--243-

 CB.gf.s54.21.f    : ->--CAG|GAC-->--122-->--TCG|GTG-->--511-->--TAG|CTC-->--124-->--TGT|TAA|AAC-->--
 annotation        : ->-----|000-->--122-->--000|----->--511-->-----|222-->--124-->--222|SSS|----->--
 prediction        : ->-----|000-->--122-->--000|----->--511-->-----|222-->--124-->--222|SSS|----->--

 CB.gf.s54.21.f    : 207-->--TCC|
 annotation        : 207-->-----|
 prediction        : 207-->-----|
-----------------------------------------------------------------------------------------------
```

Figure *5.5:* Example of a gene pair predicted **by DOUBLESCAN** which has incorrectly predicted splice sites. The splice sites are close to the annotated ones and the mis-prediction does not introduce a phase shift. The C. elegans sequence, CE.R10H10.5.f, is shown at the top, the corresponding C. ***briggsae*** sequence, CB.gf.s54.21.f, at the bottom. ***See*** Figure 4.2 for an explanation of the notation.

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

CE.Y38F1A.9.r   13003625-13008396        (4772)  reverse

 CE.Y38F1A.9.r    : |TTA-->--2928-->--AAT|TTA|GTC-->--123-->--GTT|CTG-->--912-->--GAC|ATT-->--24-->-
 annotation       : |-----<--2928--<-----|SSS|000--<--123--<--000|-----<--912--<-----|-----<--24--<-
 prediction       : |-----<--2928--<-----|SSS|000--<--123--<--000|-----<--912--<-----|000--<--24--<-

 CE.Y38F1A.9.r    : -TTT|CTG-->--198-->--TAC|CTT-->--91-->--AAC|CTA-->--44-->--TAC|CAT-->--107-->--A
 annotation       : ----|-----<--198--<-----|222--<--91--<--222|-----<--44--<-----|000--<--107--<--0
 prediction       : -000|-----<--198--<-----|222--<--91--<--222|-----<--44--<-----|000--<--107--<--0

 CE.Y38F1A.9.r    : AC|CAT|TCT-->--339-->--TTA|
 annotation       : 00|SSS|-----<--339--<-----|
 prediction       : 00|SSS|-----<--339--<-----|

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

CB.gf.s185.4.r  10405-14568     (4164)  reverse

 CB.gf.s185.4.r   : |ATA-->--2656-->--TAT|TTA|ATC-->--123-->--GTT|CTG-->--675-->--CAC|CTG-->--18-->-
 annotation       : |-----<--2656--<-----|SSS|000--<--123--<--000|-----<--675--<-----|-----<--18--<-
 prediction       : |-----<--2656--<-----|SSS|000--<--123--<--000|-----<--675--<-----|000--<--18--<-

 CB.gf.s185.4.r   : -CGG|CTC-->--89-->--TAC|CTT-->--91-->--AAC|CTG-->--97-->--CAC|CAT-->--107-->--AA
 annotation       : ----|-----<--89--<-----|222--<--91--<--222|-----<--97--<-----|000--<--107--<--00
 prediction       : -000|-----<--89--<-----|222--<--91--<--222|-----<--97--<-----|000--<--107--<--00

 CB.gf.s185.4.r   : C|CAT|TTT-->--302-->--ATC|
 annotation       : 0|SSS|-----<--302--<-----|
 prediction       : 0|SSS|-----<--302--<-----|

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

Figure *5.6:* Example of a gene pair with a wrong extra exon predicted by **DOUBLESCAN .** The extra exons are short and their length is a multiple of three base pairs thus not leading to a phase shift in the remaining correctly predicted gene structure. The C. ***elegans*** sequence, CE.Y38F1A.9.r, with the gene on the reverse strand is shown at the top, the corresponding C. *briggsae* sequence, CB.gf.s185.4.r, with the gene also on the reverse strand is shown at the bottom. See Figure 4.2 for an explanation of the notation.

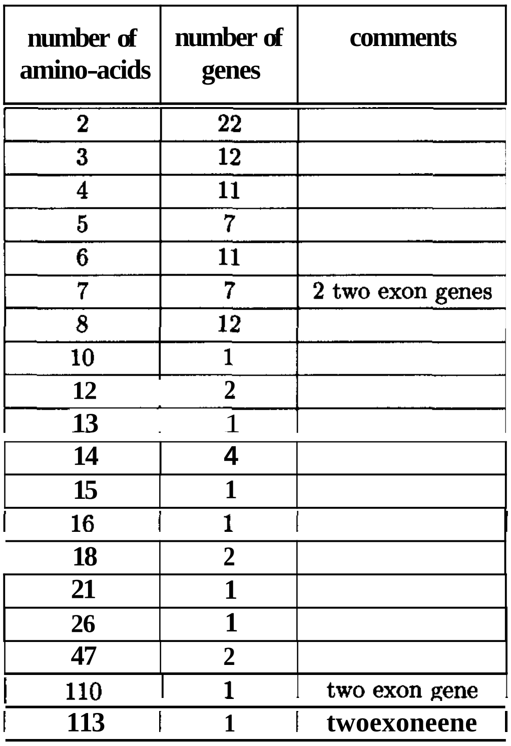
As for test set 1, the set of 268 incorrectly predicted genes which overlap an annotated gene can be subdivided into subsets according to the type of error made in the prediction, see Table 5.2 for an overview.

The dominant type of error (accounting for 27 % of errors in this test set and 26 % of errors in test set 1) consists of splitting the gene into two or more genes which overlap the annotated gene. As for test set 1, the overlap between the predicted genes and the annotated gene is very large.

As for test set 1, the incorrect or missing prediction of stop codons is another common type of error accounting for 21 % of errors (22 % in test set 1). In most cases (52 out of the 69) is a splice site predicted close to the annotated stop codon and the stop codon is missing from the prediction, see Figure 5.4. Another common type of error accounting for 22 % of errors are incorrectly predicted splice sites. This type of error is less common in test set 1 where it accounts for only 17 % of the errors. The vast majority of incorrectly predicted splice sites (64 out of 70) does not lead to a phase shift. Though the predicted splice sites are not always in close vicinity to the annotated splice sites, at least some of them may correspond to splice sites which are used in alternative splicing.

Incorrectly predicted start codons or start codons which are missing in the predicted gene account for only 11 % of the errors in this test set, whereas this type of error was more prevalent in test set 1 (accounting for 22 % of errors). Of the 21 incorrectly predicted start codons, 17 cases are mis-predictions due to a shortened or enlarged initial exon, the cases typically look like Figure 5.2 and may be due to incorrectly annotated start codons. In 14 out of the 16 genes with missing start codon, a splice site is introduced in close vicinity to the annotated start codon, see Figure 5.3 for a typical example. However, as opposed to the errors made in test set 1, the incorrect or missing prediction of the start codon leads in no case to a phase shift, i.e. the overlap between the amino-acid sequence encoded in the predicted and the annotated gene is generally high.

The remaining errors are wrong exons (accounting for 12 % of errors in this test set and for 10 % of errors in test set 1), missing introns (3 % of errors in this test set and 2 % of errors in test set 1), missing exons (3 % of errors in this test set and 1 % of errors in test set 1) and inserted introns (1 % of errors in this test set and no errors in test set 1). The majority of wrong exons (36 out of 39 cases) entail no phase shift as does none of the missing or inserted introns or missing exons.



number of amino-acids	number of genes	comments
2	22	
3	12	
4	11	
5	7	
6	11	
7	7	2 two exon genes
8	12	
10	1	
12	2	
13	1	
14	4	
15	1	
16	1	
18	2	
21	1	
26	1	
47	2	
110	1	two exon gene
113	1	twoexoneene

Table 5.4: Length distribution of the 100 wrong complete genes predicted by DOUBLESCAN on test set 2.

As opposed to test set 1 in which every annotated gene is overlapped by a predicted gene, eight genes in test set 2 are missing completely in the prediction. They correspond to four pairs of genes. The pairs of genes have the same number of exons (3, 4, 5 and 6 exons, respectively), but except for one pair of genes the lengths of the exons in one pair of genes are generally not the same. Their difference in length ranges from 3 base pairs to 18 base pairs. Overall, the four gene pairs which are missing in the prediction do not have a distinctive feature which sets them apart from the other genes.

The 10 % rate of wrong genes corresponds to 134 genes which do not overlap any annotated gene. One hundred of the 134 genes are complete genes comprising start and stop codon and the remaining 34 genes are partial genes. 82 out of the 100 complete genes encode less than ten amino-acids and almost all complete genes (96 out of 100) consist of a single exon gene. The length distribution of wrong complete genes is shown in Table 5.4. The 34 partial genes typically consist of a short initial or terminal exon comprising the start or the stop codon and a partial intron. There is no clear bias towards the 5' or 3' side of the annotated gene: 52 % of the wrong genes lie 5' to the annotated gene and 48 % 3' to the annotated gene.

Performance of the gene prediction with PROJECTOR The performance of PROJECTOR is high with a sensitivity and specificity of 90 % at gene level. The 10 % of overlapping genes corresponds to 112 genes.

About half of the incorrectly predicted genes (63 out of 112) are due to a mis-predicted splice site of one of the intermediate exons which in 51 of the 63 cases does not result in a phase shift. This type of error is much more common in this test set (51 %) than in test set 1 (31 %). As for test set 1, the incorrectly predicted splice sites are close to the annotated one and may be due to alternative splicing. The next most common source of errors are wrongly predicted intermediate exons. This type of error occurs in 28 out out the 112 incorrectly predicted genes and thus accounts for 23 % of the errors (31 % of the errors in test set 1). Incorrect start codons are only predicted in 5 of the 112 genes (corresponding to 4 %), whereas this type of error accounts for 22 % of the incorrectly predicted genes in test set 1. The mis-predicted start codon shortens or enlarges the initial exon, mostly without altering the phase within the exon. These cases may thus be due to a false annotation of the start codon rather than a false prediction by DOUBLESCAN. The rates of the other types of errors are similar to those in test set 1: in 12 % genes of the incorrectly predicted gene is an intron missing (in no case leading to a phase shift), **3** % of incorrectly predicted genes contain a wrong intron, **9** % have an incorrectly predicted stop codon and **2** % a missing exon (in all cases not leading to a phase shift). As opposed to genes with mis-predicted start codons for which the predicted and the annotated initial exons tend to have a large overlap, eight of the ten genes with an incorrectly predicted stop codon completely lack the annotated terminal exon.

5.3.3 Comparison of the performance of DOUBLESCAN and FGENESH

In order to see how well **DOUBLESCAN** does in comparison to other ab initio gene prediction programs, we compared its performance to that of **FGENESH** (Version **1.0,** nematode version of the model used with nematode parameters) [SS00]. **FGENESH** is a non-comparative ab initio gene prediction method which employs an HMM with an algorithm similar to that of **GENIE** [KHRE96] and **GENSCAN** [BK97]. **As GENSCAN, FGENESH** explicitly models the length distribution of exons and chooses its set of parameters according to the GC content of the input **DNA** sequence. Its parameters (transition and emission probabilities **as** well **as** length distributions) have been especially trained on a large set of known C. elegans genes and the underlying model has been modified to analyse nematode genes.

FGENESH is run on the two C. elegans and C. briggsae test sets (see Section **C.2** and Section **C.3** in Appendix **C)** and its performance compared to that of **DOUBLESCAN,** see Table **5.5. As** the performance for **FGENESH** is almost the same for the set of C. elegans and the set of C. briggsae genes **(as** is the case for **DOUBLESCAN),** Table **5.5** shows the performance only for the combined set of C. elegans and C. briggsae genes.

The first thing to note is that **FGENESH** has a significantly higher sensitivity and specificity on test set 1 than on test set **2.** When comparing the performance of **FGENESH** to that of **DOUBLESCAN** on test set **1, FGENESH** has a slightly higher sensitivity **(2** %) and a significantly higher specificity **(10** %) for correctly predicting entire genes. However, on test set **2 FGENESH'S** sensitivity and specificity are both significantly **(15** % and **12** %, respectively) lower than on test set **1. DOUBLESCAN'S** sensitivity is significantly higher **(7** %) than that of **FGENESH** but its specificity is again much lower **(7** %) than that of **FGENESH. For** both test sets, **DOUBLESCAN** has a very low rate of missing genes (0 % and **1** %, respectively), whereas **FGENESH** misses out **0.4** % (test set **1,** corresponding to three genes) and **11** % (test set **2,** corresponding to **112** genes) of the annotated genes completely. DOUBLESCAN's high sensitivity **for** detecting annotated genes by predicting an exactly matching **or** overlapping

	Test set 1		Test set 2	
	DOUBLESCAN	FGENESH	DOUBLESCAN	FGENESH
Gene				
Sensitivity	0.80	0.82	0.74	0.67
Specificity	0.71	0.81	0.62	0.69
Genes overlapping	0.23	0.17	0.28	0.25
Genes missing	0	0.004	0.01	0.11
Genes wrong	0.06	0.02	0.10	0.06
Start Codon				
Sensitivity	0.96	0.91	0.96	0.79
Specificity	0.87	0.94	0.81	0.86
Stop Codon				
Sensitivity	0.96	0.96	0.93	0.83
Specificity	0.89	0.96	0.82	0.88
Exon				
Feature Level				
Sensitivity	0.93	0.94	0.91	0.82
Specificity	0.90	0.93	0.89	0.87
Exons overlapping	0.06	0.04	0.07	0.05
Exons missing	0.004	0.02	0.02	0.13
Exons wrong	0.04	0.03	0.04	0.08
Nucleotide Level				
Sensitivity	0.996	0.987	0.98	0.87
Specificity	0.991	0.971	0.99	0.93

Table 5.5: Performance figures for DOUBLESCAN and FGENESH on the two *C. elegans* and *C. briggsae* test sets. The predictions by DOUBLESCAN were generated using the Stepping Stone algorithm. The table does not include the performance on the *C. elegans* and *C. briggsae* sequences separately as they are very similar. See Table 3.1 for the definitions of rows.

	Combined test set	
	DOUBLESCAN	FGENESH
Gene		
Sensitivity	0.77	0.73
Specificity	0.65	0.74
Genes overlapping	0.26	0.22
Genes missing	0.005	0.07
Genes wrong	0.09	0.04
Start Codon		
Sensitivity	0.96	0.84
Specificity	0.83	0.89
Stop Codon		
Sensitivity	0.94	0.88
Specificity	0.85	0.91
Exon		
Feature Level		
Sensitivity	0.92	0.86
Specificity	0.89	0.89
Exons overlapping	0.07	0.05
Exons missing	0.01	0.09
Exons wrong	0.04	0.06
Nucleotide Level		
Sensitivity	0.986	0.906
Specificity	0.993	0.944

Table 5.6: Performance figures for DOUBLESCAN and FGENESH on the combined *C. elegans* and *C. briggsae* test set (test set 1 and test set 2). The predictions by DOUBLESCAN were generated using the Stepping Stone algorithm. See Table 3.1 for the definitions of rows.

gene is counterbalanced by its higher rate of wrong genes (4 % higher on both test sets) with respect to FGENESH. As discussed in Section 5.3.1 and Section 5.3.2 and as shown in Table 5.3 and Table 5.4, these wrong genes are mainly short complete single exon genes which could be removed in a post-processing step.

DOUBLESCAN'S sensitivity for detecting start codons is significantly higher than that of FGENESH on both test sets (by 5 % and 17 %, respectively) and is the same for the two test sets, whereas its specificity is lower than that of FGENESH by 7 % and 5 %, respectively. For stop codons, DOUBLESCAN has almost the same sensitivity for both test sets (96 % and 93 %, respectively), whereas that of FGENESH decreases from 96 % on test set 1 to 83 % on test set 2. As for start codons, FGENESH has a higher specificity than DOUBLESCAN (7 % and 6 %, respectively), and both, DOUBLESCAN'S and FGENESH's specificity decrease from test set 1 to test set 2.

At exon level, both DOUBLESCAN and FGENESH show a high sensitivity and specificity on test set 1 with FGENESH having a 3 % higher specificity. However, on test set 2 FGENESH's sensitivity and specificity are significantly lower than on test set 1 (12 % and 6 %, respectively), whereas those of DOUBLESCAN almost stay the same (minus 2 % and minus 1 %, respectively). On test set 1, DOUBLESCAN misses almost no exons (0.4 %) and also FGENESH has a low rate of missing exons (2 %), but FGENESH's rate rises to 13 % on test set 2, whereas that of DOUBLESCAN remains low (2 %). Note that also FGENESH's rate of wrong exons changes from 3 % on test set 1 to 8 % on test set 2.

Table 5.6 shows the performance of DOUBLESCAN and FGENESH on the combined test set comprising test set 1 and 2.

5.4 Summary and discussion

Both DOUBLESCAN and PROJECTOR show very high sensitivity and specificity for predicting entire *C. elegans* and C. briggsae genes correctly and we therefore conclude that both methods, initially trained to analyse mouse and human DNA sequences, can be successfully adapted to analyse C. elegans and C. *briggsae* DNA sequences.

DOUBLESCAN has a higher sensitivity for genes, start and stop codons and exons and a significantly reduced rate of missing genes and exons compared to FGENESH, but shows a lower specificity for genes, start and stop codons. Given the fact that the training of DOUBLESCAN for C. elegans and C. briggsae involved no manual optimisation of the transition probabilities, the performance of **DOUBLESCAN** compares favorably with that of **FGENESH** and could probably be further improved.

When comparing the performances of **DOUBLESCAN** and **PROJECTOR** between the two test sets (see Table C.4 in Appendix C) and studying the sources of errors in detail (see Table 5.2), it is interesting ***to*** note that the main difference between the two test sets, namely the higher divergence of gene structures in the gene pairs of test set 2, and the difference in error rates, namely the highly increased rate of incorrectly predicted splice sites in test set **2,** may be linked.

One possible explanation is that test set **2** consists indeed of more diverged pairs of genes and that **DOUBLESCAN** and **PROJECTOR** have simply more difficulty predicting them correctly. However, another possible explanation is that the C. elegans and C. **briggsae** genes of test set 2 contain more mis-annotated splice sites and that the pairs of genes thus appear to be more diverged than they really are. This may be one of the reasons why the **BLASTN** matches covered only 95 % of the annotated exons (refer to Section **C.l** in Appendix C). In order to decide which of the two explanations holds, every gene predicted by **DOUBLESCAN** and **PROJECTOR** would have to be experimentally verified. However, one way for getting an indication **as** to which explanation is likely to be true, would be to verify whether the predicted genes are covered more by **BLASTN** hits than the annotated ones.