Appendix A

Mouse human training and test sets

A.l The training set

The training set was derived from the data set in [JBD99] which consists of pairs of orthologous mouse and human **DNA** sequences which each comprise exactly one complete gene, i.e. comprising all protein coding parts of the gene. The data set in [JBD99] was derived from the **EMBL** nucleotide database (release **55)** [SMS$^+$98] by searching human **DNA** sequences for orthologous mouse **DNA** sequences using **BLASTN** [AGM$^+$90] and by then manually inspecting the **BLASTN** results with MSPCRUNCH and **BLIXEM** [SD94]. We discarded those sequence pairs from the data set in [JBD99] which had non-consensus start or stop codons or in-frame stop codons. The remaining sequence pairs were used *to* derive the emission probabilities according to Section **2.3.** The **36** pairs of genes with consensus **GT–AG** splice sites were used to train the transition probabilities of the pair HMM by manually optimising the performance, see Section **2.3.** This data set is referred to **as** the mouse human training set.

Table **A.l** shows the basic statistics of this training set. The human and mouse genome can be divided into long GC isochores according to their GC contents and the density of genes is correlated with the GC contents. The sequences of the training set are not evenly distributed into the four GC contents intervals **as** defined by [Ber89] **as** can be seen in Table **A.2.** Within each pair, the GC contents of the two **DNA** sequences are well correlated, see Table **A.3.** Table **A.4** shows the levels of conservation of gene structures within the pairs of the training set. For the majority of pairs **(61** %), the genes in a pair have the same number of exons, but a different coding length. **36 %** of the pairs consist of evolutionarily well conserved genes which have both the same number of exons and the same coding length and only **3 %** of pairs

	min	max	mean ± standard deviation	unit
training set				
number of exons per gene	**1**	**41**	**8.3 ± 7.6**	
coding length of gene	**318**	5232	**1250 ± 964**	base pairs
length of **DNA**	**1903**	**21911**	**7256 ± 4293**	base pairs
length of gene	**1032**	**21105**	**6071 ± 4320**	base pairs
GC contents	**0.40**	**0.66**	**0.52 ± 0.06**	
test set				
number of exons per gene	**1**	**14**	**3.6 ± 2.8**	
coding length of gene	**276**	**2121**	**910 ± 477**	base pairs
length of **DNA**	**576**	**23076**	**3300 ± 2679**	base pairs
length of gene	**309**	**9033**	**2066 ± 1601**	base pairs
GC contents	**0.33**	**0.72**	**0.54 ± 0.07**	

Table **A.l:** Statistics of the mouse human training and test set. The coding length of a gene is the sum of lengths of its exons, and the length of a gene is the distance in base pairs between the start codon and the stop codon.

consist of genes which are related by events of exon-fusion or exon-splitting.

A.2 The test set

The test set was derived from the list of mouse human orthologs in [Pac99] by discarding all **DNA** pairs whose genes have non-consensus splice sites. This resulted in a set of **80** sequence pairs which is called the test set. Each **DNA** sequence in the test set comprises exactly one complete gene.

As can be seen by comparing the statistics of the training set to that **of** the test set (see Table **A.1**), the test set contains shorter genes with fewer exons in shorter **DNA** sequences. The sequences **of** the test set are more biased towards high GC contents than those in the training set, see Table **A.2. As** for the training set, **also** the GC contents of the genes within each pair of the test set are well correlated, see Table **A.3.** The test set has a higher proportion of pairs with well conserved gene structures **as 42 %** of the pairs consist of genes with the

GC contents GC contents	training set training set	test set
[0.0, 0.43)	0.06	0.05
[0.43, 0.51)	0.30	0.28
[0.51, 0.57)	0.47	0.32
[0.57, 1.00]	0.17	0.35

Table **A.2:** Distribution of GC contents in the mouse human training and test sets.

	min	max	mean ± standard deviation
training set			
mean GC contents of pair	**0.40**	**0.64**	**0.52 ± 0.05**
difference in GC contents in pair	**0.002**	**0.09**	**0.03 ± 0.02**
test set			
mean GC contents of pair	**0.38**	**0.68**	**0.54 ± 0.07**
difference in GC contents in pair	**0.00**	**0.11**	**0.04 ± 0.03**

Table **A.3:** Distribution of GC contents in the sequence pairs of the mouse human training and test sets.

	training set	test set test set
same coding length same number of exons	**0.36**	**0.42**
same coding length different number of exon	**0.00**	**0.00**
different coding length same number of exons	**0.61**	**0.55**
different coding length different number of exon	**0.03**	**0.03**

Table **A.4:** Conservation of gene structures in the gene pairs of the mouse human training and test sets.

same number of exons and the same coding length (opposed to only 36 % in the training set), see Table A.4. As for the training set, also the majority (55 %) of the test set consists of pairs in which the genes have the same number of exons, but a different coding length. Only 3 % of the gene pairs are related by events of exon-fusion or exon-splitting.

Eight genes (10 %) of the genes of the test set are also found in the training set. When removing them from the test set, the performance of Table 3.1 remains almost unchanged with most positive and negative changes within 1 % and all within 3 %.

A.3 Post-processing of the predicted mouse and human genes

In the post-processing step all predicted genes with introns of less than or equal to 50 base pairs length and or a total coding length of less than or equal to 120 base pairs length are removed.

Appendix B

Mouse human parameter tables

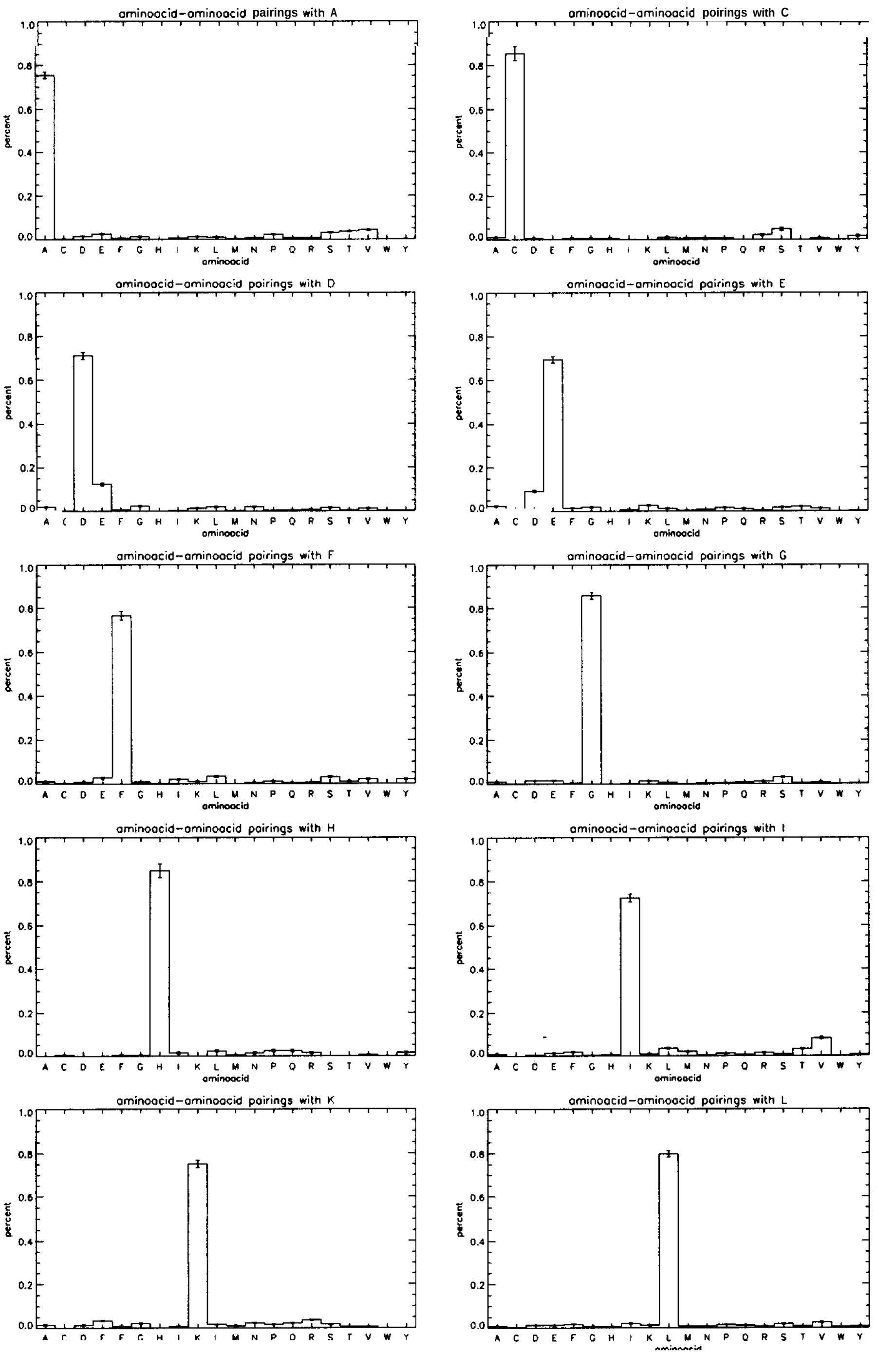
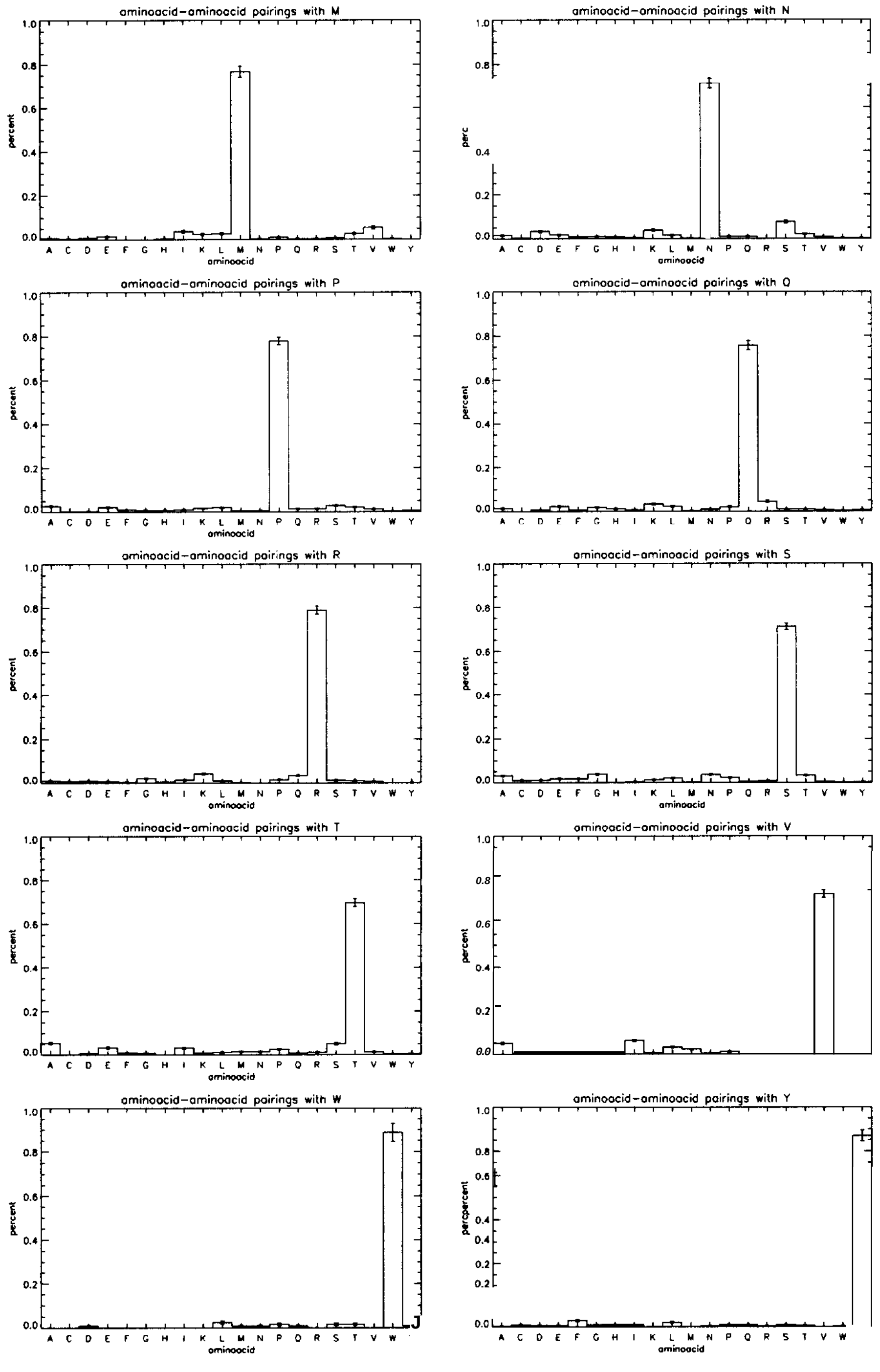


Figure B.1: Amino-aid statistics derived from the emission probabilities of the *match* exon state as determined from the training set of mouse and human DNA. The error bars indicate the statistical errors.


aminoacid–aminoacid pairings with M
aminoacid–aminoacid pairings with N
aminoacid–aminoacid pairings with P
aminoacid–aminoacid pairings with Q
aminoacid–aminoacid pairings with R
aminoacid–aminoacid pairings with S
aminoacid–aminoacid pairings with T
aminoacid–aminoacid pairings with V
aminoacid–aminoacid pairings with W
aminoacid–aminoacid pairings with Y
percent
1.0
0.8
0.6
0.4
0.2
0.0
A C D E F G H I K L M N P Q R S T V W Y
aminoacid

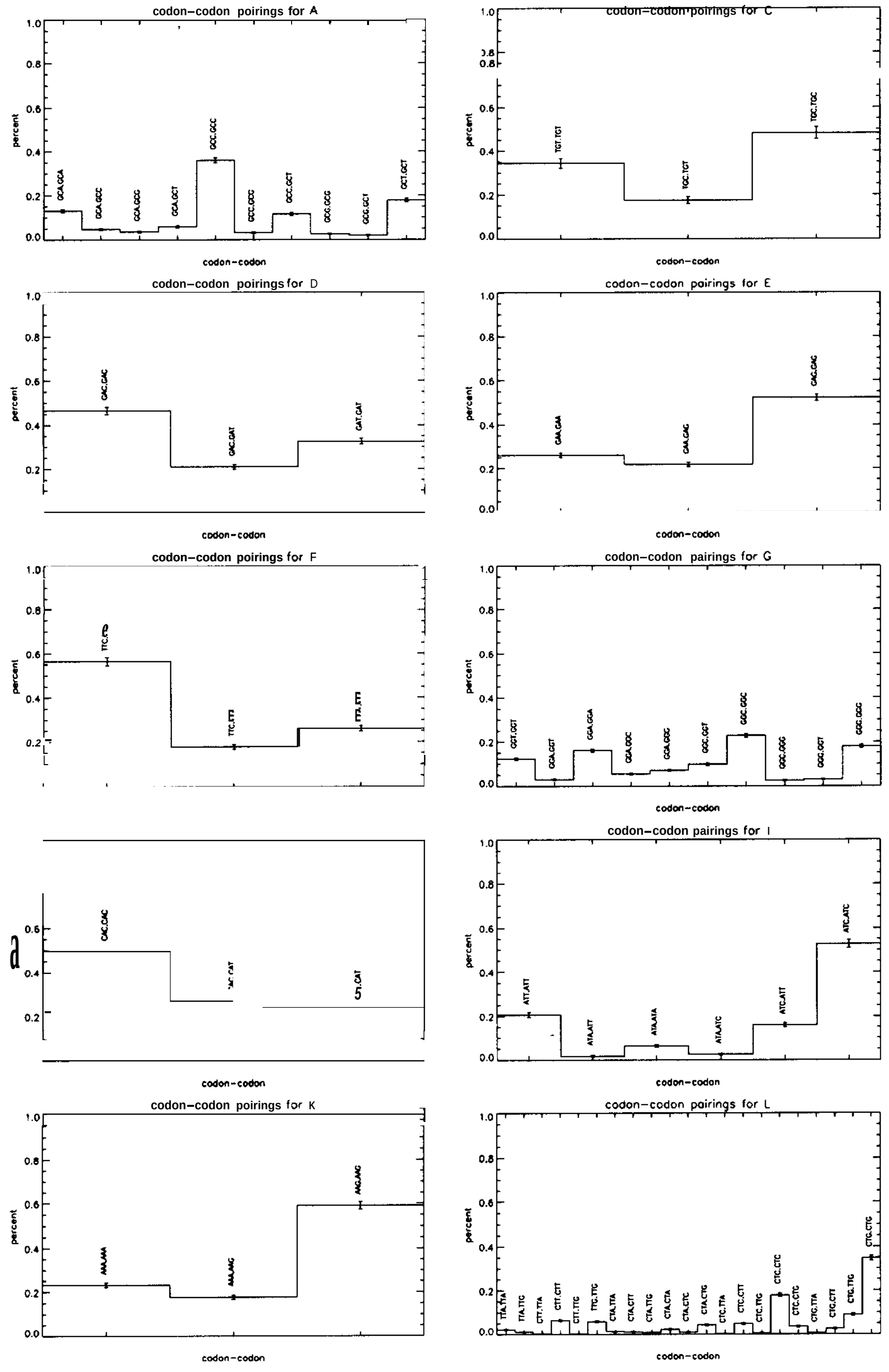
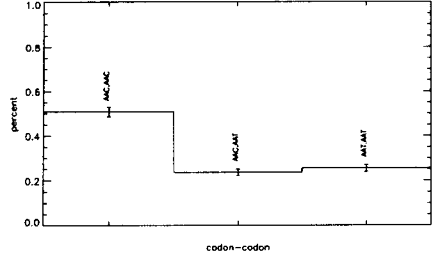
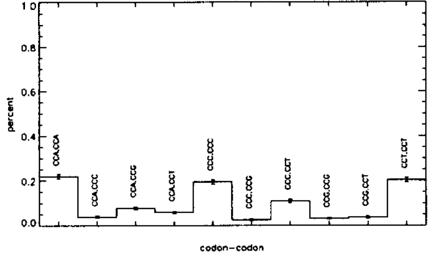


Figure **B.2:** Codon usage statistics derived from the emission probabilities of the *match exon* and the *STOP STOP* state as determined from the training set of mouse and human **DNA.** The error bars indicate the statistical errors.

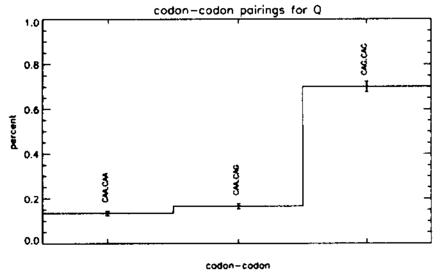
codon–codon poirings for N



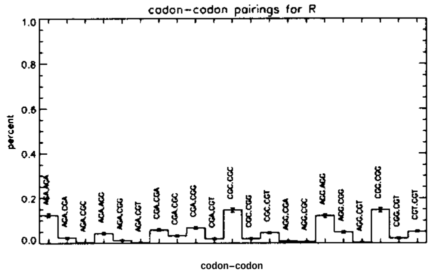
codon–codon poirings for P



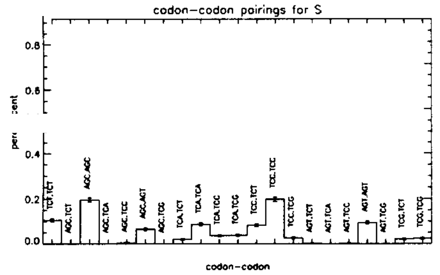
codon–codon poirings for Q



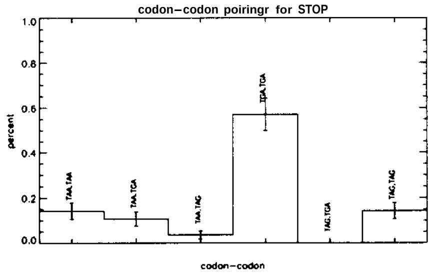
codon–codon poirings for R



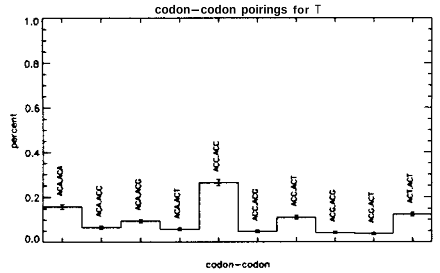
codon–codon poirings for S



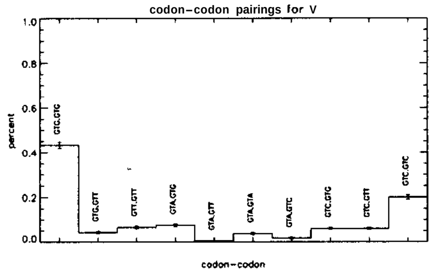
codon–codon poiringr for STOP



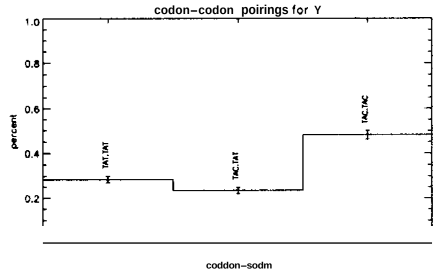
codon–codon poirings for T



codon–codon pairings for V



codon–codon poirings for Y



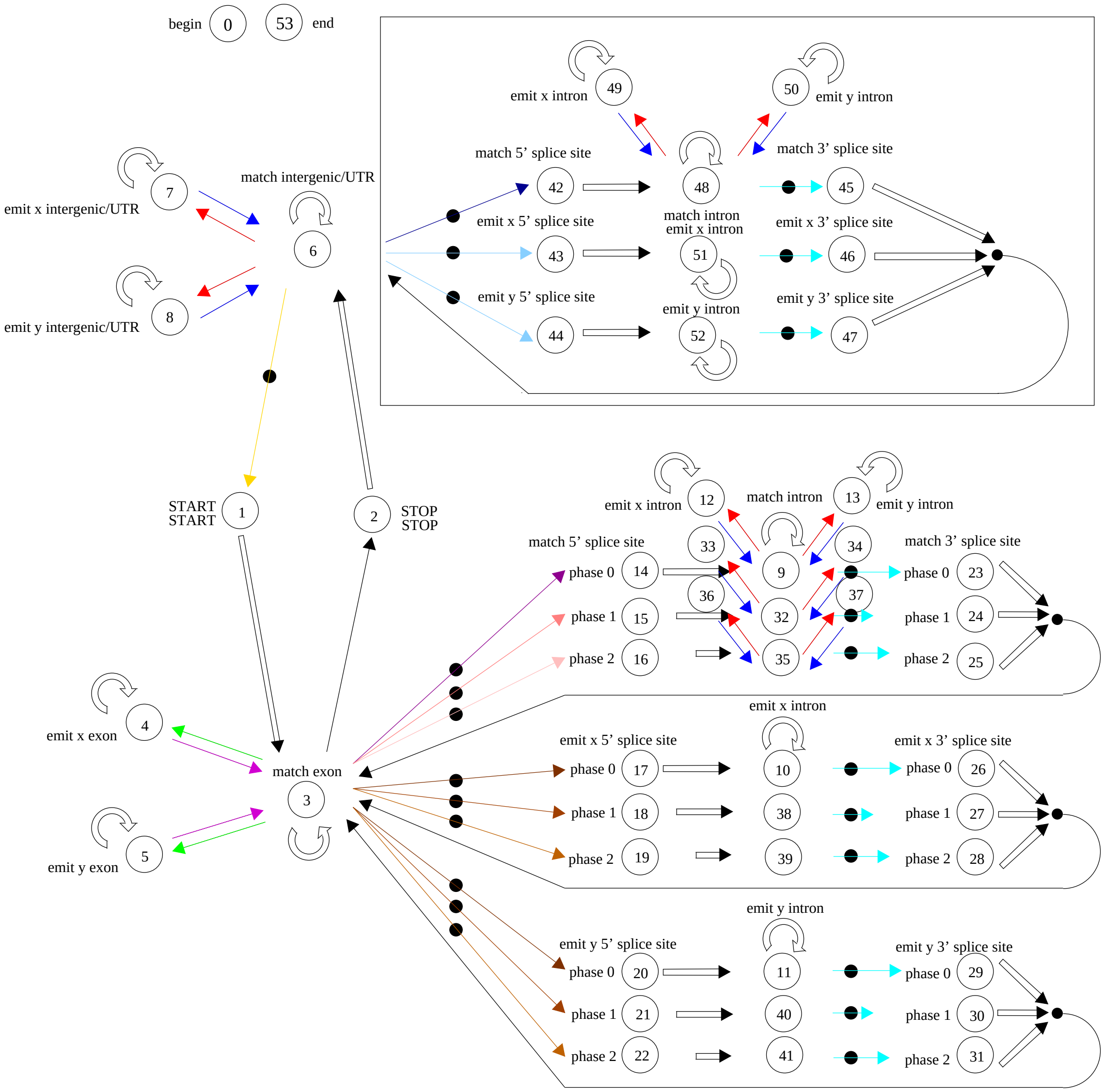


Figure B.3: States and transitions of the pair HMM underlying DOUBLESCAN and PROJECTOR. States are shown as circles, transitions as arrows. The *begin* state is connected to every state except itself and the *end* state. Likewise, there are transitions to the *end* state from every state except the *begin* state and itself. The arrows corresponding to these transitions are not shown for clarity. Each open arrow corresponds to a transition probability which is defined by the constraint that the probabilities of the transitions emerging from every state have to add up to one. Coloured arrows of the same colour correspond to transitions of the same probability. Arrows marked by a black dot are special transitions, see Section 2.3. The large box at the top right contains the states which model introns within untranslated regions (UTR-splicing).

from state	to state		derivation
match exon	*emit x exon*		(Match_exon_to_emit_exon)/2 · (1 − To_end)
	emit y exon		(Match_exon_to_emit_exon)/2 · (1 − To_end)
	STOP STOP		(Match_exon_to_stop_exon) · (1 − To_end)
	match 5' splice site phase 0	*	Special_match_exon_to_intron · Phase0 · (1 − To_end)
	match 5' splice site phase 1	*	Special_match_exon_to_intron · Phase1 · (1 − To_end)
	match 5' splice site phase 2	*	Special_match_exon_to_intron · (1 − Phase0 − Phase1)· (1 − To_end)
	emit x 5' splice site phase 0	*	Special_match_exon_to_emit_intron/2 · Phase0· (1 − To_end)
	emit x 5' splice site phase 1	*	Special_match_exon_to_emit_intron/2 · Phase1· (1 − To_end)
	emit x 5' splice site phase 2	*	Special_match_exon_to_emit_intron/2· (1 − Phase0 − Phase1) · (1 − To_end)
	emit y 5' splice site phase 0	*	Special_match_exon_to_emit_intron/2 · Phase0· (1 − To_end)
	emit y 5' splice site phase 1	*	Special_match_exon_to_emit_intron/2 · Phase1· (1 − To_end)
	emit y 5' splice site phase 2	*	Special_match_exon_to_emit_intron/2· (1 − Phase0 − Phase1) · (1 − To_end)
	match exon		(1 − Match_exon_to_stop_exon − Match_exon_to_emit_exon − Match_exon_to_match_5_splice_site − Match_exon_to_emit_5_splice_site)· (1 − To_end)
	end		To_end
match intergenic/UTR	*emit x intergenic/UTR*		Match_non_exon_to_emit_non_exon/2 · (1 − To_end)
	emit y intergenic/UTR		Match_non_exon_to_emit_non_exon/2 · (1 − To_end)
	START START	*	Special_intergenic_to_start_exon · (1 − To_end)
	match 5' splice site	*	Special_match_exon_to_intron· 1/(Special_match_exon_to_intron + Special_match_exon_to_emit_intron)· (1 − To_end)
	emit x 5' splice site	*	Special_match_exon_to_emit_intron/2· 1/(Special_match_exon_to_intron + Special_match_exon_to_emit_intron)· (1 − To_end)
	emit y 5' splice site	*	Special_match_exon_to_emit_intron/2· 1/(Special_match_exon_to_intron + Special_match_exon_to_emit_intron)· (1 − To_end)
	match intergenic/UTR		(1 − Match_intergenic_to_start_exon − Match_non_exon_to_emit_non_exon − Match_exon_to_match_5_splice_site − Match_exon_to_emit_5_splice_site)· (1 − To_end)
	end		To_end

Table B.1: Parametrisation of the transition probabilities within the pair HMM underlying DOUBLESCAN and PROJECTOR. The values of the parameters are given in Table B.2. $N = 54$ is the number of states in the pair HMM of DOUBLESCAN and PROJECTOR. Special transitions (see Section 6.2 for details) are indicated by an asterisk (*) in the third column. Note that the nominal values of the transitions emerging from a state do not have to add up to one if one or more of the transitions are special.

from state	to state		derivation
match intron	match intron		(1 − Match_non_exon_to_emit_non_exon − Match_intron_to_match_exon) . (1 − To-end)
			same for states 9, 32, 35, 48
	emit x intron		Match_non_exon_to_emit_non_exon/2 · (1 − To_end)
			same for transitions 9 to 12, 32 to 33, 35 to 36, 48 to 49
	emit y intron		Match_non_exon_to_emit_non_exon/2 · (1 − To_end)
			same for transitions 9 to 13, 32 to 34, 35 to 37, 48 to 50
	match 3' splice site	*	Special_intron_to_match_exon · (1 − To_end)
			same for transitions 9 to 23, 32 to 24, 35 to 25, 48 to 45
	end		To_end
			same for transitions 9 to 53, 32 to 53, 35 to 53, 48 to 53
emit x exon	*match exon*		Emit_exon_to_match_exon · (1 − To_end)
	emit x exon		(1 − Emit_exon_to_match_exon) · (1 − To_end)
	end		To_end
emit y exon	*match exon*		Emit_exon_to_match_exon · (1 − To_end)
	emit y exon		(1 − Emit_exon_to_match_exon) · (1 − To_end)
	end		To_end
emit x intergenic/UTR	*match intergenic/UTR*		Emit_non_exon_to_match_non_exon · (1 − To_end)
	emit x intergenic/UTR		(1 − Emit_non_exon_to_match_non_exon) · (1 − To_end)
	end		To_end
emit y intergenic/UTR	*match intergenic/UTR*		Emit_non_exon_to_match_non_exon · (1 − To_end)
	emit y intergenic/UTR		(1 − Emit_non_exon_to_match_non_exon) · (1 − To_end)
	end		To_end
emit x intron	*emit x 3' splice site*	*	Special_intron_to_match_exon · (1 − To_end)
			same for transitions 10 to 26, 38 to 27, 39 to 28, 51 to 46
	emit x intron		(1 − Match_intron_to_match_exon) · (1 − To_end)
			same for states 10, 38, 39, 51
	end		To_end
			same for transitions 10 to 53, 38 to 53, 39 to 53, 51 to 53
emit y intron	*emit y 3' splice site*	*	Special_intron_to_match_exon · (1 − To_end)
			same for transitions 11 to 29, 40 to 30, 41 to 31, 52 to 47
	emit y intron		(1 − Match_intron_to_match_exon) · (1 − To_end)
			same for states 11, 40, 41, 52
	end		To_end
			same for transitions 11 to 53, 40 to 53, 41 to 53, 52 to 53
match 5' splice site	*match intron*		(1 − To_end)
			same for transitions 14 to 9, 15 to 32, 16 to 35, 42 to 48
	end		To_end
			same for transitions 14 to 53, 15 to 53, 16 to 53, 42 to 53
emit x 5' splice site	*emit x intron*		(1 − To_end)
			same for transitions 17 to 10, 18 to 38, 19 to 39, 43 to 51
	end		To_end
			same for transitions 17 to 53, 18 to 53, 19 to 53, 43 to 53
emit y 5' splice site	*emit y intron*		(1 − To_end)
			same for transitions 20 to 11, 21 to 40, 22 to 41, 44 to 52
	end		To_end
			same for transitions 20 to 53, 21 to 53, 22 to 53, 44 to 53

from state	to state		derivation
begin	anv connected state		**1/(*N* − 2)**
START START	match exon		**1** −To-end
	end		To_end
STOP STOP	match *intergenic/UTR*		**1** − To-end
	end		To-end
match 3' splice site	match exon or match *intergenic/UTR*		(1 − To-end)
			same for transitions **23** to **3, 24** to **3, 25** to **3, 45** to **6**
	end		To-end
			same for transitions **23** to **53, 24** to **53, 25** to **53, 45** to **53**
emit x 3' splice site	match exon or match *intergenic/UTR*		**(1** − To-end)
			same for transitions **26** to **3, 27** to **3, 28** to **3, 46** to **6**
	end		To-end
			same for transitions **26** to **53, 27** to **53, 28** to **53, 46** to **53**
emit y 3' splice site	match exon		**(1** − To-end)
			same for transitions 29 to **3, 30** to **3, 31** to **3, 47** to **6**
	end		To-end
			same for transitions 29 to **53, 30** to **53, 31** to **53, 47** to 53
emit x intron **of match intron**	match intron		**Emit_non_exon_to_match_non_exon** · (1 − To-end)
			same for transitions **12** to **9, 33** to **32, 36** to **35, 49** to **48**
	emit x intron of match intron		(1 − Emit_non_exon_to_match_non_exon) **. (1** − To-end)
			same for states **12, 33, 36, 49**
	end		To-end
			same for transitions **12** to **53, 33** to **53, 36** to **53, 49** to **53**
emit y intron ***of match intron***	match intron		**Emit_non_exon_to_match_non_exon .** (**1** − To-end)
			same for transitions **13** to **9, 34** to **32, 37** to **35, 50** to **48**
	emit y intron of match intron		**(1** − Emit_non_exon_to_match_non_exon) **. (1** − To-end)
			same for states **13, 34, 37, 50**
	end		**To_end**
			same for transitions 13 to 53, 34 to 53, 37 to 53, 50 to 53

parameter	value
Phase0	0.4387
Phase1	0.387
To-end	0.0001
Match-exon-tostop-exon	0.003
Match-exon-to-emit_exon	0.02
Match_exon_to_match_5_splice_site	5e-06
Match-exon-to-emit_5_splice_site	5e-06
Matchintergenic-tostart-exon	0.0001
Matchnon-exon-to-emitnon-exon	0.08
Matchintron-tomatch-exon	1e-05
Emit_exon_to_match_exon	0.33333
Emitnon-exon-tomatchnon-exon	0.04
Special_match_exon_to_intron	1
Special_intron_to_match_exon	0.25
Specialmatch_exon_to_emit intron	0.06666
Special_intergenic_to_start_exon	0.1

Table B.2: Values of the parameters on which the transition probabilities depend.

parameter	value
Prior-GT	0.01
Prior-GC	0.0001
PriorAG	0.001
PriorATG	0.005

Table **B.3:** Values of the priors which are used with the special transition probabilities of the pair HMM underlying **DOUBLESCAN** and **PROJECTOR** **for** the analysis of mouse and human **DNA** sequences.

Appendix *C*

C. elegans C. briggsae training and test sets

The training set of C. elegans and C. briggsae gene pairs has been established by Avril Coghlan, Trinity College, Dublin.

C.l The training set

As described in Chapter 5, the training set was used only to derive the emission probabilities of DOUBLESCAN according to Section 2.3. In particular, it was not used to derive the values of the transition probabilities nor to fine-tune the performance, see Section 5.2. The test set comprises 910 pairs of C. elegans and C. briggsae DNA sequences, each comprising exactly one complete gene. The C. elegans genes are known genes of Wormbase release WS77 [SSD$^+$01, Wor] and the *C. briggsae* genes are putative genes predicted by GENEFINDER [eSC98]. All pairs of genes were defined as being orthologous using BLAST [AGM$^+$90]. The exons of the two genes were mutual best hits and hit each other with an Evalue a hundred times smaller than the second best hit and with an E-value of less than 0.1. Pairs of orthologous exons were covered by at least 95 % by BLAST hits. Only 16 out of 910 gene pairs (1.7 % of the training set) had splice sites which were not equal to the GT–AG consensus. Table C.l shows some statistics of the training set. As opposed to the mouse and human genome which can be partitioned into long GC isochores according to their GC contents, the GC density within the C. elegans and C. briggsae genomes is uniform around 36 %, see Table C.2. However, as can be seen by comparing Table C.3 and Table A.3 in Appendix A, the GC contents of orthologous C. elegans and C. briggsae genes are **as** well correlated **as** those of orthologous mouse and human genes.

The gene structures of orthologous C. elegans and C. briggsae genes are more conserved than those of the mouse human training set (see Table A.4 in Appendix A) **as** can be seen from Table C.4. The majority **(53** %) of genes has the same exon number and coding length **as** its orthologous partner in the other genome and differences in the gene structures between orthologous genes are only due to a difference in coding length, but not in exon number.

C.2 Test set 1

As the training set is only used to automatically derive the emission probabilities of the *match* exon and *STOP STOP* state, but not for the derivation of the transition probabilities nor the fine-tuning of the performance, we can use the same data **as** a test set. Test set 1 is **a** subset of the training set. It comprises **353** pairs of genes whose exons were entirely covered by **BLAST** hits (100 %) and which either have the consensus splice sites **GT-AG** or the non-consensus splice sites **GC-AG** (present in **3** out of **353** gene pairs). The statistics *can* be found in Table C.l. Genes in this test set are on average shorter than those of test set 2 and have fewer **exons.** The orthologous genes in this test set have better conserved gene structures and are thus more closely related than those of test set 2, *see* Table C.4.

C.3 Test set 2

Also test set 2 **is** a subset of the training set. It comprises **535** pairs of genes whose exons were covered by at least **95 %** but less than 100 % by **BLAST** matches and which either have the consensus splice sites **GT-AG** or the non-consensus splice sites **GC-AG** (present in **8** out of **535** gene pairs). There is no intersection between test set 1 and test set 2. The statistics *can* be found in Table C.l. Table C.4 shows the level of conservation between the gene structures of orthologous genes.

	min	max	mean ± standard deviation	unit
training set				
number of exons per gene	1	21	4.1 ± 2.1	
coding length of gene	150	5046	917 ± 606	base pairs
length of **DNA**	461	36529	3455 ± 2818	base pairs
length of gene	180	11594	1536 ± 1187	base pairs
GC contents	0.27	0.55	0.38 ± 0.04	
test set 1				
number of exons per gene	1	13	3.5 ± 1.7	
coding length of gene	150	2988	697 ± 435	base pairs
length of **DNA**	461	19253	2994 ± 2477	base pairs
length of gene	180	7759	1191 ± 930	base pairs
GC contents	0.27	0.51	0.38 ± 0.04	
test set 2				
number of exons per gene	1	21	4.5 ± 2.3	
coding length of gene	177	5046	1058 ± 665	base pairs
length of **DNA**	560	36529	3741 ± 2988	base pairs
length of gene	225	11594	1753 ± 1286	base pairs
GC contents	0.29	0.55	0.38 ± 0.04	

Table C.l: Statistics of the C. *elegans* C. *briggsae* training and test sets. The coding length of a gene **is** the sum of lengths of its exons and the length of a gene is the distance in base pairs between the start codon and the stop codon.

GC contents	training set	test set 1	test set 2
[0.0, 0.43)	0.923	0.91	0.933
[0.43, 0.51)	0.074	0.09	0.062
[0.51, 0.57)	0.003	0.00	0.005
[0.57, 1.00]	0.000	0.00	0.000

Table **C.2:** Distribution of GC contents in the *C. elegans* C. *briggsae* training and test sets.

	min	max	mean ± standard deviation
training set			
mean GC contents of pair	0.31	0.53	0.38 ± 0.03
difference in GC contents in **pair**	0.00	0.20	0.03 ± 0.02
test set 1			
mean GC contents **of** pair	0.32	0.50	0.38 ± 0.03
difference in GC contents in **pair**	0.00	0.11	0.03 ± 0.02
test set 2			
mean GC contents of pair	0.31	0.53	0.38 ± 0.03
difference in GC contents in **pair**	0.00	0.20	0.03 ± 0.02

Table **C.3:** Distribution of GC contents in the sequence pairs **of** the C. ***elegans*** C. ***briggsae*** training and test sets.

	training set	test set 1	test set 2
same coding length same number of exons	0.53	0.997	0.21
same coding length different number of exon	0.00	0.000	0.00
different coding length same number of exons	0.47	0.003	0.79
different coding length different number of exon	0.00	0.000	0.00

Table C.4: Conservation of gene structures in the gene pairs of the C. *elegans* C. *briggsae* training and test sets.

Appendix D

C. elegans C. briggsae parameter tables

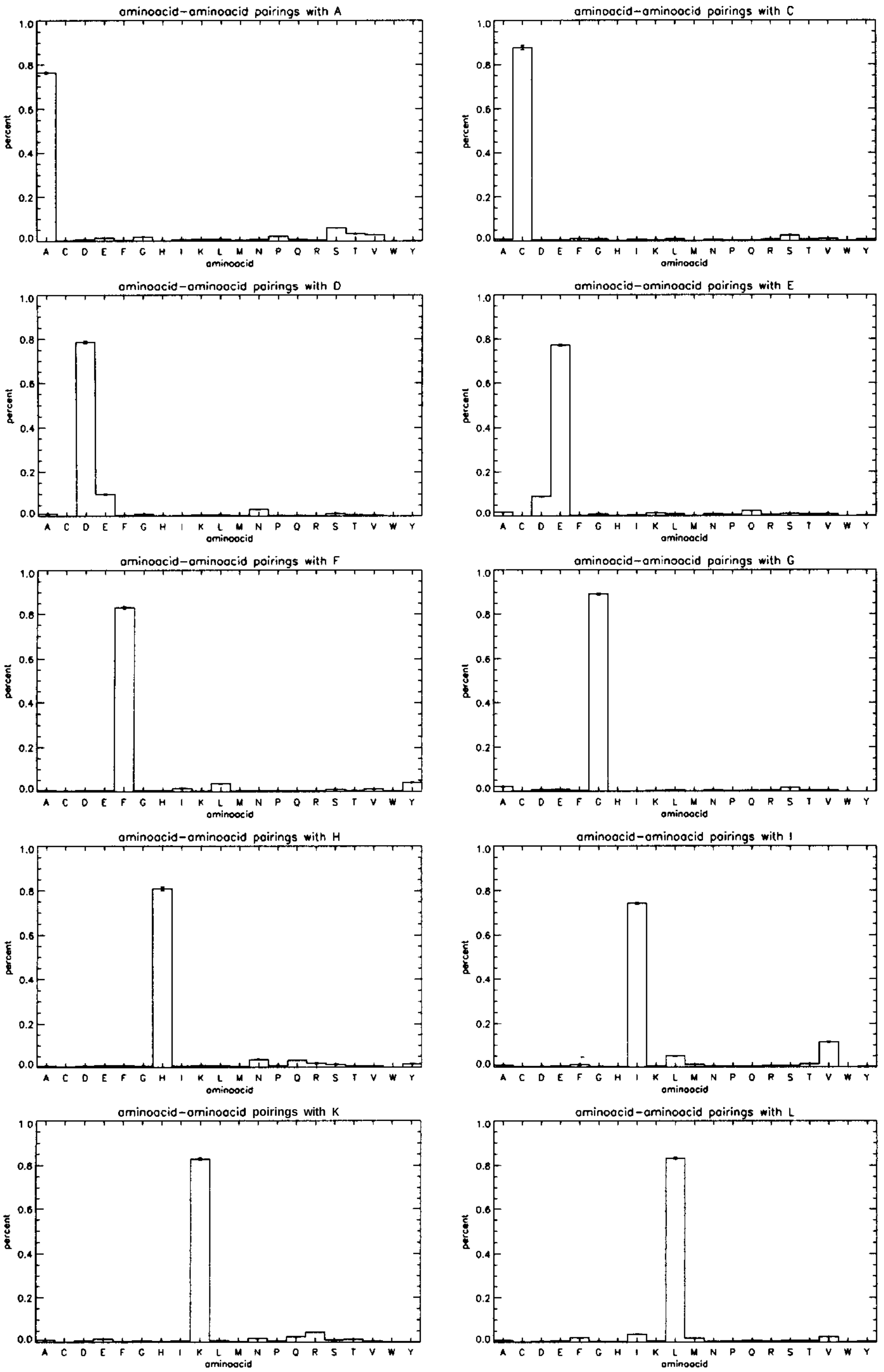
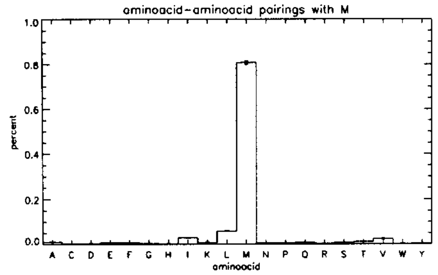
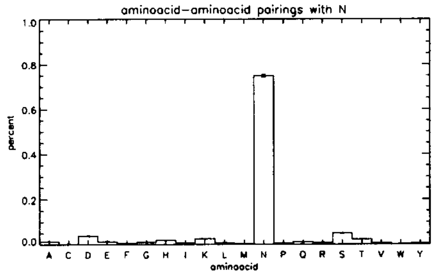
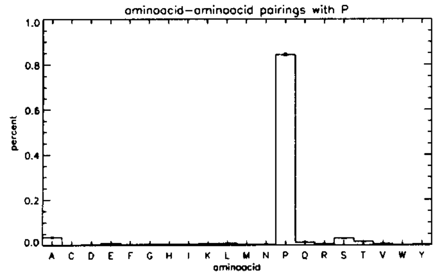
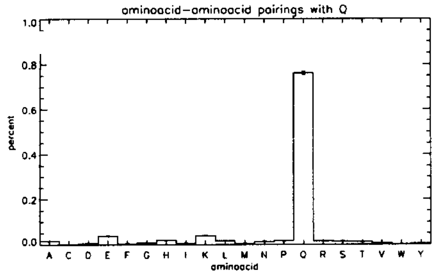


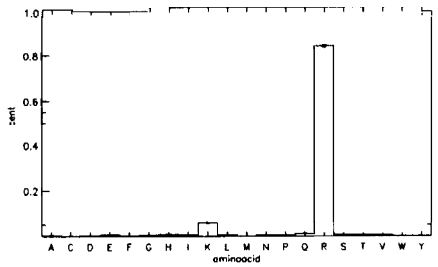
Figure D.1: **Amino-acid statistics derived from the emission probabilities** of **the match *exon* state as determined** from **the training set** of **C. *elegans* and *C. briggsae* DNA. The error** bars **indicate the statistical errors.**

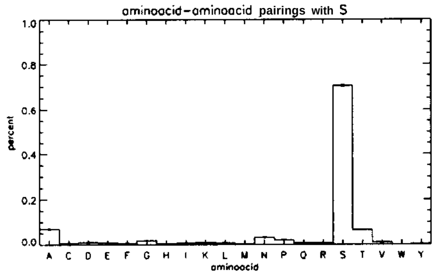

aminoacid-aminoacid pairings with M
1.0
0.8
0.6
0.4
0.2
0.0
percent
A C D E F G H I K L M N P Q R S T V W Y
aminoacid

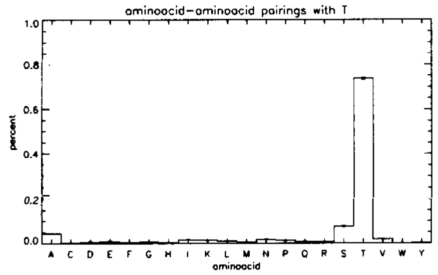

aminoacid-aminoacid pairings with N
1.0
0.8
0.6
0.4
0.2
0.0
percent
A C D E F G H I K L M N P Q R S T V W Y
aminoacid

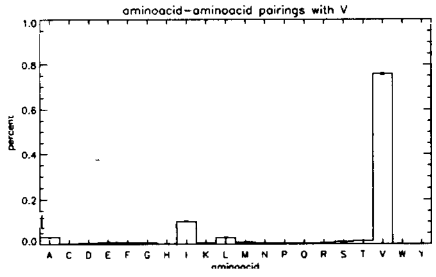

aminoacid-aminoacid pairings with P
1.0
0.8
0.6
0.4
0.2
0.0
percent
A C D E F G H I K L M N P Q R S T V W Y
aminoacid

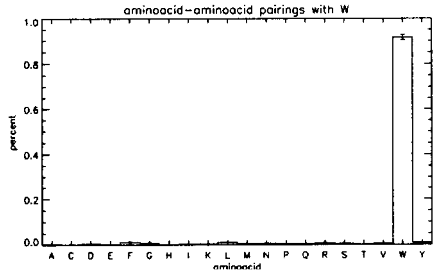

aminoacid-aminoacid pairings with Q
1.0
0.8
0.6
0.4
0.2
0.0
percent
A C D E F G H I K L M N P Q R S T V W Y
aminoacid

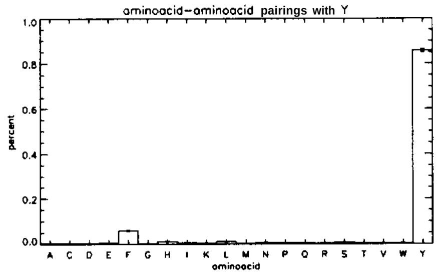

1.0
0.8
0.6
0.4
0.2
cent
A C D E F G H I K L M N P Q R S T V W Y
aminoacid


aminoacid-aminoacid pairings with S
1.0
0.8
0.6
0.4
0.2
0.0
percent
A C D E F G H I K L M N P Q R S T V W Y
aminoacid


aminoacid-aminoacid pairings with T
1.0
0.8
0.6
0.4
0.2
0.0
percent
A C D E F G H I K L M N P Q R S T V W Y
aminoacid


aminoacid-aminoacid pairings with V
1.0
0.8
0.6
0.4
0.2
0.0
percent
A C D E F G H I K L M N P Q R S T V W Y
aminoacid


aminoacid-aminoacid pairings with W
1.0
0.8
0.6
0.4
0.2
0.0
percent
A C D E F G H I K L M N P Q R S T V W Y
aminoacid


aminoacid-aminoacid pairings with Y
1.0
0.8
0.6
0.4
0.2
0.0
percent
A C D E F G H I K L M N P Q R S T V W Y
aminoacid

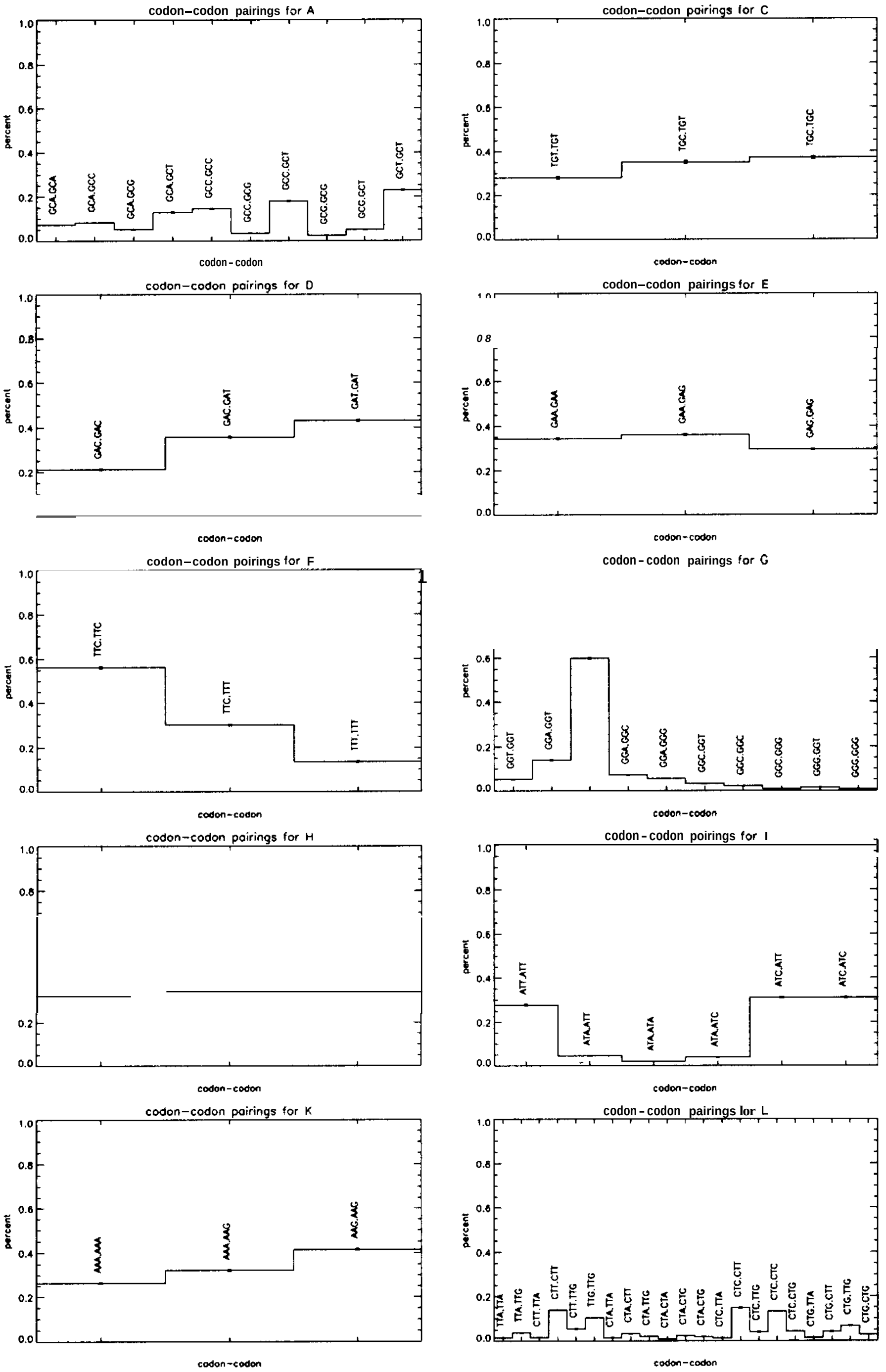
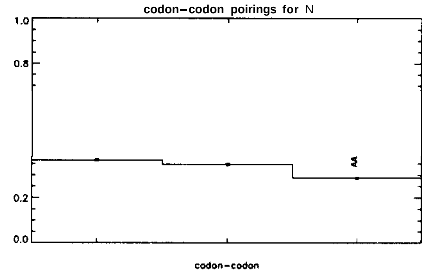
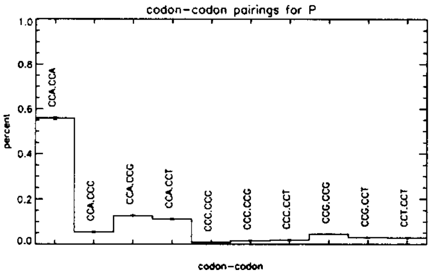
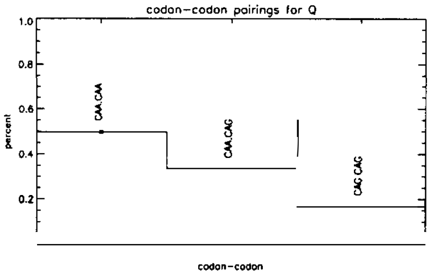
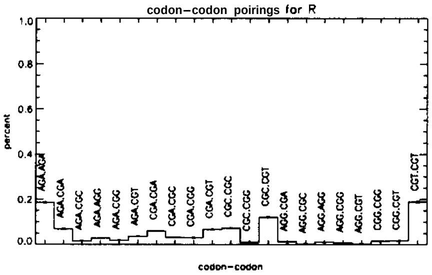


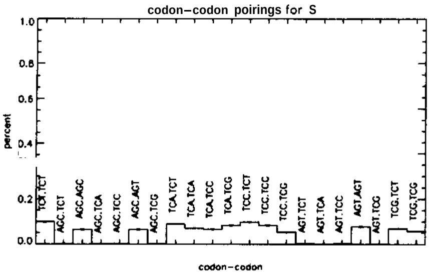
Figure D.2: Codon usage statistics derived from the emission probabilities of **the** *match* ***exon*** **and the** *STOP STOP* **state as determined from the training set** of ***C.*** elegans **and *C.*** briggsae **DNA. The error bars indicate the statistical errors.**

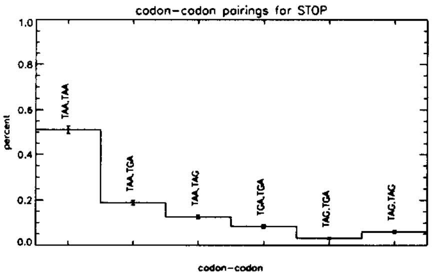

codon−codon pairings for N
1.0
0.8
0.2
0.0
AA
codon−codon

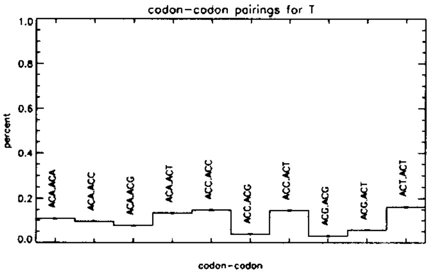

codon−codon pairings for P
1.0
0.8
0.6
0.4
0.2
0.0
percent
CCA.CCA
CCA.CCC
CCA.CCG
CCA.CCT
CCC.CCC
CCC.CCG
CCC.CCT
CCG.CCG
CCG.CCT
CCT.CCT
codon−codon

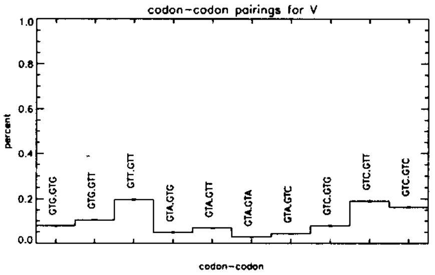

codon−codon pairings for Q
1.0
0.8
0.6
0.4
0.2
percent
CAA.CAA
CAA.CAG
CAG.CAG
codon−codon

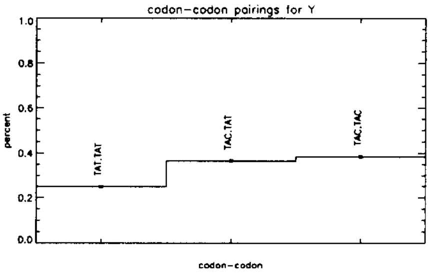

codon−codon poirings for R
1.0
0.8
0.6
0.4
0.2
0.0
percent
AGA.AGA
AGA.CGA
AGA.CGC
AGA.AGG
AGA.CGG
AGA.CGT
CGA.CGA
CGA.CGC
CGA.CGG
CGA.CGT
CGC.CGC
CGC.CGG
CGC.CGT
AGG.CGA
AGG.CGC
AGG.AGG
AGG.CGG
AGG.CGT
CGG.CGG
CGG.CGT
CGT.CGT
codon−codon


codon−codon poirings for S
1.0
0.8
0.6
0.4
0.2
0.0
percent
TCT.TCT
AGC.TCT
AGC.AGC
AGC.TCA
AGC.TCC
AGC.AGT
AGC.TCG
TCA.TCT
TCA.TCA
TCA.TCC
TCA.TCG
TCC.TCT
TCC.TCC
TCC.TCG
AGT.TCT
AGT.TCA
AGT.TCC
AGT.AGT
AGT.TCG
TCG.TCT
TCG.TCG
codon−codon


codon−codon pairings for STOP
1.0
0.8
0.6
0.4
0.2
0.0
percent
TAA.TAA
TAA.TGA
TAA.TAG
TGA.TGA
TAG.TGA
TAG.TAG
codon−codon


codon−codon pairings for T
1.0
0.8
0.6
0.4
0.2
0.0
percent
ACA.ACA
ACA.ACC
ACA.ACG
ACA.ACT
ACC.ACC
ACC.ACG
ACC.ACT
ACG.ACG
ACG.ACT
ACT.ACT
codon−codon


codon−codon pairings for V
1.0
0.8
0.6
0.4
0.2
0.0
percent
GTG.GTG
GTG.GTT
GTT.GTT
GTA.GTG
GTA.GTT
GTA.GTA
GTA.GTC
GTC.GTG
GTC.GTT
GTC.GTC
codon−codon


codon−codon pairings for Y
1.0
0.8
0.6
0.4
0.2
0.0
percent
TAT.TAT
TAC.TAT
TAC.TAC
codon−codon

parameter	value	comment
Prior_GT	0.01	
Prior_GC	0.0001	
Prior_AG	0.01	PriorAG (mousehuman) = 0.001
Prior_ATG	0.005	

Table D.1: Values of the priors which are used with the special transition probabilities of the pair HMM underlying DOUBLESCAN and PROJECTOR for the analysis of C. *elegans C.* briggsae DNA sequences. The value of the prior for the 3' splice sites (PriorAG) is the *only* transition parameter which is different from the parameters used for the analysis of mouse and human DNA (see Table B.3 in Appendix B). The parametrisation of the transition probabilities as well as the values of the parameters for the analysis of C. *elegans C. briggsae* DNA sequences are the same as those for the analysis of mouse human DNA sequences (see Table B.1 and Table B.2 in Appendix B).

Appendix E

The DOUBLESCAN web-server

DOUBLESCAN can be accessed via a web-server at

www.sanger.ac.uk/Software/analysis/doublescan/

DOUBLESCAN needs as input two DNA sequences in a variant of the FASTA format which requires a modified header-line:

>name start_position–end_position orientation

(see also www.sanger.ac.uk/Software/analysis/doublescan/fasta_format.shtml) where

- name is the name of the sequence (example: Mm)
- start-position is an integer which is the position of the first character in the sequence (example: **100)** and its value has to be smaller to that of the end-position
- end-position is an integer which is the position of the last character in the sequence (example: **737** i.e. the sequence is **737-100+1** = **638** nucleotides long)
- orientation can be either 'forward' or 'reverse' depending on the strand which is to be analysed for genes. Note that the value of the orientation in the header line does not indicate the orientation of the sequence as the FASTA file should always give the sequence of the forward strand.
- the fields in the header line have to be tab-delimited

To give an example of an input file in the required Fasta format:

```
>Mm 100-737 forward
gggaatgaagtttttctgcaggatttaaatgtggtctttaagagacaccgcatgcaaaga
atagctggggcttgctagccaatgaaaacattcagattccaatgacgcatcctttttttct
ccaccccccttccaagacccggattcggaaaccccgcctaacgctctagttttcaaccagg
tccgcagaaggcctatttaaagggacgattgctgtctccctgctgtcataaccatgtctg
gacgtggcaagggtggtaaaggccttgggaaaggcggcgctaagcgccaccgtaaggttc
tccgcgataacatccagggcatcaccaagcctgccatccgccgcctggcccggcgcgggg
gagtgaagcgcatctccggcctcatctacgaggagacccgcggtgtgctgaaggtgttcc
tggagaacgtgatccgcgacgccgtcacctacacggagcacgccaagcgcaagaccgtca
ccgccatggacgtggtctacgcgctcaagcgccagggccgcactctctacggattcggcg
gttaatcgactaacaaacgattttccactgtcaacaaaaggcccttttcagggccaccca
caaattcctagaaggagttgttcacttaccgaagctt
```

Every analysis by Doublescan returns two output files:

- a file containing the predicted annotation of the two input DNA sequences in gtf format (see http://www.fruitfly.org/flyannot/format.html#GTF)
- a file containing the predicted annotation of the two input DNA sequences and the predicted conserved subsequences in a variant of the gtf format

The following example shows an output file in gtf-format which indicates the predicted annotation:

```
Mm  Doublescan  Start_Codon  234  236  .  +  0  gene_id  3;  transcript_id  3;  exon_number  1
Mm  Doublescan  CDS          234  542  .  +  0  gene_id  3;  transcript_id  3;  exon_number  1
Mm  Doublescan  Stop_Codon   543  545  .  +  0  gene_id  3;  transcript_id  3;  exon_number  1
Mm  Doublescan  Exon         234  545  .  +  .  gene_id  3;  transcript_id  3;  exon_number  1

Hs  Doublescan  Start_Codon  311  313  .  +  0  gene_id  7;  transcript_id  7;  exon_number  1
Hs  Doublescan  CDS          311  619  .  +  0  gene_id  7;  transcript_id  7;  exon_number  1
Hs  Doublescan  Stop_Codon   620  622  .  +  0  gene_id  7;  transcript_id  7;  exon_number  1
Hs  Doublescan  Exon         311  622  .  +  .  gene_id  7;  transcript_id  7;  exon_number  1
```

The corresponding output file in the modified gtf-format indicates the predicted annotation as well as the conserved subsequences:

```
Mm  Doublescan  Intergenic     1   61  .  +  .  conserved
Mm  Doublescan  Intergenic    62  103  .  +  .
Mm  Doublescan  Intergenic   104  133  .  +  .  conserved
Mm  Doublescan  Intergenic   134  154  .  +  .
Mm  Doublescan  Intergenic   155  187  .  +  .  conserved
Mm  Doublescan  Intergenic   188  209  .  +  .
Mm  Doublescan  Intergenic   210  233  .  +  .  conserved
Mm  Doublescan  Start_Codon  234  236  .  +  0  gene_id  3;  transcript_id  3;  exon_number  1;  conserved
Mm  Doublescan  CDS          237  542  .  +  0  gene_id  3;  transcript_id  3;  exon_number  1;  conserved
Mm  Doublescan  Stop_Codon   543  545  .  +  0  gene_id  3;  transcript_id  3;  exon_number  1;  conserved
Mm  Doublescan  Intergenic   546  551  .  +  .  conserved
Mm  Doublescan  Intergenic   552  560  .  +  .
Mm  Doublescan  Intergenic   561  568  .  +  .  conserved
Mm  Doublescan  Intergenic   569  571  .  +  .
Mm  Doublescan  Intergenic   572  609  .  +  .  conserved
Mm  Doublescan  Intergenic   610  631  .  +  .
Mm  Doublescan  Intergenic   632  637  .  +  .  conserved

Hs  Doublescan  Intergenic     1   26  .  +  .
Hs  Doublescan  Intergenic    27   74  .  +  .  conserved
Hs  Doublescan  Intergenic    75  115  .  +  .
Hs  Doublescan  Intergenic   116  180  .  +  .  conserved
Hs  Doublescan  Intergenic   181  216  .  +  .
Hs  Doublescan  Intergenic   217  248  .  +  .  conserved
Hs  Doublescan  Intergenic   249  307  .  +  .
Hs  Doublescan  Intergenic   308  310  .  +  .  conserved
Hs  Doublescan  Start_Codon  311  313  .  +  0  gene_id  7;  transcript_id  7;  exon_number  1;  conserved
Hs  Doublescan  CDS          314  619  .  +  0  gene_id  7;  transcript_id  7;  exon_number  1;  conserved
Hs  Doublescan  Stop_Codon   620  622  .  +  0  gene_id  7;  transcript_id  7;  exon_number  1;  conserved
Hs  Doublescan  Intergenic   623  665  .  +  .  conserved
Hs  Doublescan  Intergenic   666  844  .  +  .
Hs  Doublescan  Intergenic   845  859  .  +  .  conserved
```