Appendix A . Tables related to the investigation of tRNA-gene order conservation in mammalian genomes

This appendix contains tables related to the investigation of tRNA-gene order conservation in the mammalian genomes (see chapter 2, section 2.2)

TGC	Ala1	GTG	His1	TGG	Pro1	GCT	Ser5
GGC	Ala2	ATG	His2	GGG	Pro2	ACT	Ser6
CGC	Ala3	TAT	Ile1	CGG	Pro3	TGT	Thr1
AGC	Ala4	GAT	Ile2	AGG	Pro4	GGT	Thr2
GCA	Cys1	AAT	Ile3	TTG	Gln1	CGT	Thr3
ACA	Cys2	TTT	Lys1	CTG	Gln2	AGT	Thr4
GTC	Asp1	CTT	Lys2	TCG	Arg1	TAC	Val1
ATC	Asp2	TAA	Leu1	GCG	Arg2	GAC	Val2
TTC	Glu1	CAA	Leu2	CCG	Arg3	CAC	Val3
CTC	Glu2	TAG	Leu3	ACG	Arg4	AAC	Val4
GAA	Phe1	GAG	Leu4	TCT	Arg5	CCA	Trp1
AAA	Phe2	CAG	Leu5	CCT	Arg6	GTA	Tyr1
TCC	Gly1	AAG	Leu6	TGA	Ser1	ATA	Tyr2
GCC	Gly2	CAT	Met1	GGA	Ser2	TTA	Ter1
CCC	Gly3	GTT	Asn1	CGA	Ser3	CTA	Ter2
ACC	Gly4	ATT	Asn2	AGA	Ser4	TCA	Sec1

Table A 1. Lookup table of anticodon types and the tRNA-gene symbols

cluster ID	chr	start	end
1.1.10	1	16,719,667	17,088,832
2.1.2	1	93,754,422	94,085,801
3.1.42	1	142,481,551	148,284,076
4.1.36	1	159,636,114	159,858,162
5.1.2	1	165,950,586	165,951,420
6.1.3	1	202,742,278	203,709,966
7.1.2	1	247,134,677	247,135,141
8.2.2	2	27,127,154	27,127,658
9.2.2	2	130,749,494	130,811,242
10.2.2	2	156,965,527	156,965,975
11.3.2	3	133,430,634	133,433,403
12.3.2	3	149,703,918	149,799,324
13.5.17	5	180,456,676	180,582,073
14.6.150	6	26,394,733	29,064,839
15.6.8	6	58,249,836	58,304,654
16.6.2	6	144,579,377	145,545,623
17.7.20	7	148,638,214	149,035,764
18.8.4	8	66,772,086	67,189,050
19.11.8	11	59,074,678	59,090,501

cluster ID	chr	start	end
20.11.2	11	75,624,205	75,624,588
21.12.2	12	97,421,412	97,422,232
22.12.5	12	123,972,254	123,990,536
23.13.2	13	40,532,874	40,928,132
24.14.14	14	20,147,335	20,222,086
25.15.3	15	43,278,096	43,280,712
26.15.2	15	76,939,959	77,824,124
27.16.17	16	3,140,676	3,359,885
28.16.2	16	22,114,533	22,216,043
29.16.2	16	55,891,364	55,891,975
30.16.5	16	69,369,615	70,017,969
31.17.18	17	7,963,198	8,071,107
32.17.2	17	19,352,086	19,704,837
33.17.8	17	34,161,560	35,527,152
34.17.3	17	70,541,596	70,542,875
35.18.2	18	41,553,749	41,923,341
36.19.2	19	1,334,361	1,334,635
37.19.2	19	4,675,082	4,675,719
38.X.3	X	3,766,418	3,843,344

Table A 2 The start and end coordinates of the tRNA gene clusters in the human genome (assembly NCBI 36)

Each cluster identifier (ID) is composed of three numbers separated by ".". The first number is a serial number. The second number (or X) is the chromosome on which a particular cluster resides. The third number is the number of tRNA gene loci in a particular cluster.

chr: chromosome

cluster ID	chr	start	end
1.1.3	1	73,971,393	74,985,840
2.1.2	1	107,331,203	107,332,257
3.1.2	1	134,861,508	134,861,945
4.1.2	1	167,478,309	167,479,017
5.1.26	1	172,870,617	173,506,186
6.2.2	2	56,997,464	56,997,850
7.2.2	2	118,738,191	118,747,667
8.2.3	2	122,066,935	122,069,480
9.3.2	3	3,109,391	3,135,216
10.3.5	3	19,820,110	20,371,715
11.3.2	3	51,446,283	51,447,407
12.3.30	3	96,396,659	97,766,935
13.4.2	4	56,953,853	57,727,180
14.4.3	4	131,397,335	132,386,642
15.4.2	4	149,499,077	150,476,012
16.5.2	5	31,164,664	31,165,168
17.5.5	5	125,693,626	125,698,919
18.5.2	5	142,649,903	142,755,501
19.6.52	6	47,908,583	48,294,102
20.6.2	6	86,211,030	86,369,597
21.7.2	7	28,081,759	28,502,820
22.7.3	7	98,690,607	99,418,054
23.7.2	7	120,626,628	120,708,747
24.8.2	8	97,592,760	97,593,215

cluster ID	chr	start	end
25.8.3	8	113,517,230	113,949,306
26.9.2	9	64,181,087	64,536,123
27.9.3	9	104,258,153	104,266,736
28.10.3	10	61,786,481	62,824,914
29.10.2	10	79,652,093	79,652,361
30.10.2	10	90,611,211	90,611,967
31.11.8	11	48,661,965	48,700,478
32.11.2	11	58,118,372	58,118,775
33.11.18	11	68,853,198	68,941,443
34.11.2	11	94,705,047	95,675,333
35.11.6	11	97,518,539	97,805,084
36.11.3	11	115,229,071	115,229,941
37.12.2	12	16,346,839	16,877,619
38.13.60	13	21,168,250	22,058,232
39.13.46	13	23,277,886	23,618,045
40.14.7	14	49,985,834	50,012,669
41.16.2	16	3,012,435	3,364,711
42.17.8	17	23,261,584	23,277,957
43.17.2	17	35,195,954	35,288,056
44.19.2	19	3,066,129	3,576,335
45.19.8	19	12,069,281	12,079,383
46.X.2	X	13,016,125	13,859,646
47.X.15	X	131,542,096	131,936,800
48.X.2	X	156,110,215	156,479,321

Table A 3. The start and end coordinates of the tRNA gene clusters in the mouse genome (assembly NCBI M36).

The convention used to assign the cluster ID to each cluster is the same as that used in Table A 2

human clusters (NCBI36)	mouse clusters (NCBI M36)	conservation type	quality of the human genome assembly	qualitiy of the mouse genome assembly
1.1.10	NA	synteny-non-conserved	CSN	FCS
2.1.2	coord: 3.122284970.122285054.-1	single-conserved	FCS	FCS
3.1.42	12.3.30	complicated	CSN	FCS
4.1.36	5.1.26	gapped	FCS	FCS
5.1.2	4.1.2	perfect	FCS	FCS
6.1.3	3.1.2	sub perfect type two	FCS	FCS
7.1.2	32.11.2	perfect	FCS	FCS
8.2.2	16.5.2	perfect	FCS	FCS
9.2.2	coord: 1.34379358.34379429.-1	single-conserved	FCS	FCS
10.2.2	6.2.2	perfect	FCS	FCS
11.3.2	27.9.3	sub perfect type two	FCS	FCS[1]
12.3.2	NA	synteny-non-conserved	FCS	WGS
13.5.17	31.11.8	gapped	FCS	FCS
14.6.150	38.13.60/39.13.46	gapped	FCS	FCS
15.6.8	NA	synteny-non-conserved	CSN	FCS
16.6.2	coord: 10.12612761.12612843.-1	single-conserved	FCS	FCS
17.7.20	19.6.52	gapped	FCS	FCS
18.8.4	10.3.5	sub perfect type two	FCS	FCS
19.11.8	45.19.8	sub perfect type one	FCS	FCS
20.11.2	22.7.3	sub perfect type two	FCS	FCS
21.12.2	30.10.2	perfect	FCS	FCS
22.12.5	17.5.5	sub perfect type one	FCS	FCS
23.13.2	NA	synteny-non-conserved	FCS	FCS
24.14.14	40.14.7	gapped	FCS	FCS
25.15.3	8.2.3	sub perfect type one	FCS	FCS
26.15.2	coord: 9.89924402.89924474.-1	single-conserved	FCS	FCS
27.16.17	42.17.8	gapped	FCS	FCS[2]
28.16.2	23.7.2	sub perfect type one	FCS	FCS
29.16.2	24.8.2	perfect	FCS	FCS
30.16.5	25.8.3	gapped	FCS	FCS[3]
31.17.18	33.11.18	sub perfect type one	FCS	FCS

human clusters (NCBI36)	mouse clusters (NCBI M36)	conservation type	quality of the human genome assembly	qualitiy of the mouse genome assembly
32.17.2	coord: 11.61224111.61224182.-1	single-conserved	FCS	FCS
33.17.8	35.11.6	gapped	FCS	FCS
34.17.3	36.11.3	perfect	FCS	FCS
35.18.2	NA	synteny-non-conserved	FCS	CSN
36.19.2	29.10.2	perfect	FCS	FCS
37.19.2	NA	synteny-non-conserved	FCS	FCS
38.X.3	NA	synteny-non-conserved	CSN	WGS

Table A 4. The synteny conservation of clustered human tRNA gene loci in the mouse genome

For columns 1 and 2 the cluster IDs are taken from Table A 1 and Table A 2 for human and mouse respectively.

NA: not available (when there is no corresponding cluster in the mouse genome).

coord: “coordinate” of a singlet tRNA gene locus in the mouse genome. This is used when the syntenic counterpart in the mouse genome is a singlet. The convention used here is chromosome:start:end:strand.

FCS: finished contig sequence; CSN: unfinished contig sequence (with gaps); WGS: whole genome shotgun sequence

[1]: mouse WGS between the 3' end tRNA gene and 3' boundary of the syntenic block

[2]: mouse WGS in the upstream region of the 5' end tRNA gene in this cluster

[3]: mouse WGS between the 5' end tRNA gene and 5' boundary of the syntenic block

singlet ID	coordinate (NCBI36)	coordinate (NCBIM36)	quality of the human genome assembly	quality of the mouse genome assembly
nc.1	1.55196130.55196202.-1	NA	FCS	FCS
nc.2	1.151910350.151910421.1	3.90561787.90561858.-1	FCS	FCS
nc.3	1.157378025.157378098.-1	1.175227004.175227077.1	FCS	FCS
nc.4	1.170424162.170424230.-1	NA	FCS	FCS
nc.5	1.178450899.178450971.-1	NA	FCS	FCS
nc.6	1.220704970.220705042.1	NA	FCS	CSN
nc.7	2.42891180.42891272.1	17.83770270.83770362.1	FCS	FCS
nc.8	2.70329627.70329697.-1	6.86369527.86369597.1	FCS	FCS
nc.9	2.74977554.74977622.1	NA	FCS	FCS
nc.10	2.117498979.117499050.-1	NA	FCS	FCS
nc.11	2.218818794.218818886.1	NA	FCS	FCS
nc.12	3.45705495.45705567.-1	9.123378123.123378195.-1	FCS	FCS
nc.13	3.126895867.126895938.-1	NA	FCS	FCS
nc.14	3.170972712.170972784.1	3.30792108.30792180.1	FCS	FCS
nc.15	3.185848789.185848859.-1	NA	FCS	FCS
nc.16	4.40603500.40603572.-1	NA	FCS	FCS
nc.17	4.124649455.124649526.-1	NA	FCS	FCS
nc.18	4.156604428.156604502.-1	NA	FCS	FCS
nc.19	5.26234296.26234368.-1	NA	FCS	FCS
nc.20	5.141754172.141754243.-1	NA	FCS	FCS
nc.21	5.159324619.159324696.-1	NA	FCS	FCS
nc.22	6.18944381.18944452.1	NA	FCS	WGS
nc.23	6.37395973.37396045.1	NA	FCS	FCS
nc.24	6.69971099.69971181.1	NA	FCS	FCS
nc.25	6.126143086.126143157.-1	10.30500556.30500627.1	FCS	FCS
nc.26	6.142620469.142620539.1	NA	FCS	FCS
nc.27	7.98905243.98905314.1	NA	FCS	FCS
nc.28	7.128210740.128210811.1	6.29338834.29338905.1	FCS	FCS
nc.29	7.138675986.138676058.1	6.38463539.38463611.1	FCS	FCS
nc.30	8.59667352.59667422.1	NA	FCS	FCS
nc.31	8.96351061.96351142.-1	4.10801211.10801292.1	FCS	FCS
nc.32	8.124238651.124238723.-1	15.57806066.57806138.-1	FCS	FCS
nc.33	9.5085085.5085156.1	NA	FCS	FCS
nc.34	9.14423938.14424009.-1	4.82090854.82090925.-1	FCS	FCS
nc.35	9.19393996.19394070.1	NA	FCS	FCS

singlet ID	coordinate (NCBI36)	coordinate (NCBIM36)	quality of the human genome assembly	quality of the mouse genome assembly
nc.36	9.76707810.76707881.-1	NA	FCS	FCS
nc.37	9.112000624.112000696.1	NA	FCS	FCS
nc.38	9.114656810.114656908.1	NA	FCS	FCS
nc.39	9.125695343.125695415.-1	NA	FCS	FCS
nc.40	9.130142176.130142266.-1	NA	FCS	FCS
nc.41	10.5935680.5935752.-1	NA	WGS	FCS
nc.42	10.22558444.22558517.-1	2.18504798.18504871.-1	WGS	FCS
nc.43	10.69194267.69194348.1	10.62824833.62824914.-1	FCS	FCS
nc.44	11.9253366.9253439.1	NA	FCS	FCS
nc.45	11.45246776.45246849.-1	NA	FCS	FCS
nc.46	11.50190455.50190526.-1	NA	FCS	FCS
nc.47	11.51216476.51216548.1	NA	FCS	FCS
nc.48	11.65872167.65872248.1	19.5038304.5038385.-1	FCS	FCS
nc.49	11.108541249.108541330.1	NA	FCS	FCS
nc.50	11.121935865.121935937.1	NA	FCS	FCS
nc.51	12.27734573.27734645.1	NA	FCS	FCS
nc.52	12.54870415.54870496.1	10.127861413.127861494.-1	FCS	FCS
nc.53	12.73137449.73137521.1	NA	FCS	FCS
nc.54	12.94953930.94954001.1	10.92882777.92882848.-1	FCS	FCS
nc.55	12.121426877.121426947.1	NA	FCS	FCS
nc.56	13.30146101.30146174.-1	5.149539350.149539423.-1	FCS	WGS
nc.57	13.44390062.44390133.-1	14.74886929.74887000.1	FCS	FCS
nc.58	13.93999905.93999977.-1	14.116971871.116971943.-1	FCS	FCS
nc.59	14.22468750.22468822.1	14.53471901.53471973.1	FCS	FCS
nc.60	14.31306567.31306637.-1	NA	FCS	FCS
nc.61	14.57776366.57776438.-1	12.71887725.71887797.-1	FCS	FCS
nc.62	14.72499432.72499503.1	NA	FCS	FCS
nc.63	14.88515195.88515267.1	NA	FCS	FCS
nc.64	14.101853182.101853255.1	12.111293145.111293218.1	FCS	FCS
nc.65	15.23878474.23878545.-1	7.58267184.58267255.1	FCS	FCS
nc.66	15.38673315.38673396.-1	2.118738191.118738272.-1	FCS	FCS
nc.67	15.63948454.63948525.-1	9.64536052.64536123.1	FCS	FCS
nc.68	15.87679308.87679380.1	7.79339932.79340004.1	FCS	FCS
nc.69	16.626737.626807.1	17.25602688.25602758.1	FCS	FCS
nc.70	16.14287251.14287322.1	16.13350901.13350972.1	FCS	FCS

singlet ID	coordinate (NCBI36)	coordinate (NCBIM36)	quality of the human genome assembly	quality of the mouse genome assembly
nc.71	16.72069717.72069789.-1	NA	FCS	FCS
nc.72	16.85975129.85975201.-1	8.124465281.124465353.-1	FCS	FCS
nc.73	17.15349410.15349483.1	NA	FCS	FCS
nc.74	17.26901213.26901284.1	11.79520845.79520916.1	FCS	FCS
nc.75	17.44624889.44624960.1	11.95675262.95675333.-1	FCS	FCS
nc.76	17.56218375.56218445.1	NA	CSN	FCS
nc.77	17.59957380.59957453.-1	NA	FCS	FCS
nc.78	17.63446475.63446547.-1	11.106828956.106829028.1	FCS	FCS
nc.79	17.78045886.78045957.-1	NA	CSN	FCS
nc.80	19.19713207.19713277.1	NA	FCS	WGS
nc.81	19.38359803.38359876.1	7.34943530.34943603.-1	FCS	FCS
nc.82	19.40758590.40758662.1	NA	FCS	FCS
nc.83	19.44594648.44594740.-1	7.28081759.28081853.1	FCS	FCS
nc.84	19.50673700.50673785.-1	7.18459766.18459851.1	FCS	FCS
nc.85	19.54729745.54729817.-1	NA	FCS	FCS
nc.86	19.57117208.57117280.-1	NA	FCS	WGS
nc.87	20.17803142.17803219.1	NA	FCS	FCS
nc.88	20.48385749.48385830.-1	NA	FCS	FCS
nc.89	21.14848387.14848457.1	NA	FCS	FCS
nc.90	21.17748978.17749048.-1	NA	FCS	FCS
nc.91	22.42877870.42877955.1	NA	FCS	FCS
nc.92	X.18602950.18603022.-1	X.156110215.156110287.1	FCS	FCS

Table A 5. The synteny conservation of non-clustered human tRNA gene loci (singlets) in the mouse genome

NA: not available (when there is no corresponding cluster in the mouse genome)

The assignment of a singlet ID follows the convention: “nc” (non-clustered). “serial number”.

The coordinates presented here follow the convention of that used in Table A 4.

Appendix B. The program sets written for this thesis

This appendix lists the main program sets that were particularly written for this thesis

Program set 1:

Table B 1. Functions of the program sets written for this thesis

Name: Search for synteny-conserved ncRNAs
Description of function:
Search for synteny-conserved ncRNAs in syntenic regions between two genomes, and determine the number of covariations between each pair of orthologous ncRNAs that are synteny-conserved. For each ncRNA locus in a particular genome, this program set can search for its corresponding syntenic blocks, which are defined by the unique best reciprocal homologue pairs (UBRHPs) that are determined by Ensembl, in other genome(s). For a particular ncRNA in one genome, its synteny-conserved counterpart is searched for in the corresponding syntenic region of the other genome initially using WUBLAST. This blast hit is then structurally aligned, using cmsearch (a program in the Infernal package) (Griffiths-Jones et al. 2003), to its consensus RNA structure, and the number of covariations between each pair of orthologous ncRNAs that are synteny-conserved are determined.

Program set 2

Name: Search and process synteny-conserved tRNA-gene Cluster
Description of function:
Search for synteny-conserved tRNA-gene clusters in the syntenic regions between two genomes, and examine the gene-order difference between two orthologous tRNA-gene clusters. For a tRNA-gene cluster in the human genome, this program set can search for its corresponding synteny-conserved clusters in other genomes in the syntenic regions defined by UBRHPs. A pair of orthologous tRNA-gene clusters are further analyzed by comparing the gene-order conservation between them.

Program set 3

Name: Align two ordered list of (tRNA-)gene symbols
Description of function:
Examine the gene-order conservation between two lists of (tRNA-)gene symbols. Using the dynamic programming library functions provided by biojava, this program set can align two lists of tRNA-gene symbols, which may be derived from a pair of syntenic regions from two genomes.

Program set 4

Name: RNA folding package
Description of function:
Predict the RNA secondary structure of a given sequence, and report the locations and sizes of stems and loops in this sequence. This program set provides an implementation of the Zuker's RNA secondary structure predicting algorithm. The thermodynamic parameters follow the ones used in (Zuker 1989). A set of adjunctive functions are provided in this program set, in order to facilitate the retrieval of local hairpins and the calculation of their thermodynamic stabilities.

Program set 5

Name: Eponine RNA motif extension, anchored
Description of function:
Prepare local hairpins and perform training of an Eponine anchored model which may consist of a set of RNA motifs. This program set provides a mechanism to extend Eponine anchored models to model RNA motifs. For each sequence recruited for training an Eponine anchored model, local RNA structures are predicted for each windowed region using Zuker's RNA secondary-structure predicting algorithm. Then SimpleStemLoopBasisSource uses the parameters of local hairpins as the basis to propose a new model. Other classes with the suffix BasisSource can optimize the parameters of a model using Monte Carlo sampling approaches. The parameters of an anchored model containing RNA motifs may consist of distributions of hairpin dimensions and/or stability and distance distributions between each motif and the anchored point of each sequence.

Program set 6

Name: Eponine RNA motif extension, unanchored
Description of function:
Prepare local hairpins and perform the learning of an Eponine unanchored model which may consist of a set of RNA motifs. This program set provides a mechanism to extend the Eponine unanchored models to model RNA motifs. For each sequence recruited for training an Eponine unanchored model, local RNA structures are predicted for each windowed region in this sequence using Zuker's algorithm. Then ConvolvedSensorsBasis uses the parameters of local hairpins as the basis to propose a new model. Other classes with the suffix BasisSource can optimize the parameters of a model by using Monte Carlo sampling approaches. The parameters of an unanchored model containing RNA motifs may consist of distributions of hairpin dimensions and/or stability and distance distributions between motifs.

Table B 2. Number of lines and file sizes of the program sets written for this thesis

Chap	Program set	Number of Lines	File size
2	Search-for-synteny-conserved-ncRNAs		
	● syntenic_proteins.pm	870	36k
	● protein_boundary.pm	320	9k
	● infernal.pm	192	5k
	● best_blast_hit.pm	103	6k
	● cmsearch_hit.pm	103	2k
	● paired_cmsearch_hit.pm	486	13k
	● other miscellaneous modules and scripts	1580	40k
2	Search-process-synteny-conserved-tRNACluster		
	● tRNAClusterDB.pm	130	3k
	● tRNASeqFasta.pm	321	3k
	● tRNAInfo.pm	112	2k
	● tRNAClusterDB_protein_boundary.pl	889	16k
	● other miscellaneous modules and scripts	274	8k
2, 3	Align two ordered list of (tRNA-)gene symbols		
	● AligntRNAName.java	511	15k
	● Other miscellaneous classes	209	7k
4, 6	RNA folding package		
	● Stem.java	32	1k
	● AbstractStem.java	52	2k
	● SimpleStem.java	204	5k
	● StemTools.java	93	3k
	● StrucTools.java	632	19k
	● StrucReport.java	183	4k
	● Pair.java	60	2k
	● Zuker.java	822	23k
4, 6	Eponine RNA motif extension, anchored		

	● AbstractStructureSampler.java	168	5k
	● SimpleStemLoopConstraint.java	793	23k
	● SimpleStemLoopBasisSource.java	348	9k
	● LocalEnergyDistBasisSource.java	59	2k
	● LocalEnergyOffsetBasisSource.java	59	2k
	● LoopSizeDistBasisSource.java	71	2k
	● LoopSizeOffsetBasisSource.java	71	2k
	● StemEnergyDistBasisSource.java	59	2k
	● StemEnergyOffsetBasisSource.java	59	2k
	● StemSizeDistBasisSource.java	69	2k
	● StemSizeOffsetBasisSource.java	71	2k
4, 6	Eponine RNA motif extension, unanchored		
	● AbstractStrucSampler.java	223	6k
	● ConvolvedSensorsBasis.java	798	24k
	● NewStruc1.java	380	10k
	● SampleLocalEnergyDist.java	65	2k
	● SampleLocalEnergyOffset.java	65	2k
	● SampleLoopSizeDist.java	71	2k
	● SampleLoopSizeOffset.java	76	2k
	● SampleStemEnergyDist.java	65	2k
	● SampleStemEnergyOffset.java	65	2k
	● SampleStemSizeDist.java	74	2k
	● SampleStemSizeOffset.java	75	2k