

Using phylogenetics and model selection
to investigate the evolution of RNA
genes in genomic alignments

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF LIFE SCIENCES

2013

James Edward Allen

Contents

List of Tables	7
List of Figures	9
Abstract	10
Declaration	11
Copyright	12
Acknowledgements	13
1 Introduction	14
1.1 RNA genes	15
1.1.1 Types of non-coding RNA	16
1.1.2 RNA secondary structure	18
1.1.3 RNA gene prediction	21
1.2 Sequence alignment	22
1.2.1 Multiple sequence alignment and phylogenetics	23
1.2.2 Genomic alignment	23
1.3 Phylogenetic inference	24
1.3.1 Distance matrices	25
1.3.2 Parsimony	25
1.3.3 Maximum likelihood and Bayesian analysis	25
1.3.4 Choosing an inference method	26
1.4 Substitution models of evolution	28
1.4.1 DNA models	28
1.4.2 RNA models	29
1.4.3 Amino acid models	30

1.4.4	Rate heterogeneity	31
1.4.5	Codon models	32
1.4.6	Context-dependent models	32
1.4.7	Model assumptions	33
1.4.8	Model selection	33
1.5	Aims and objectives	34
1.5.1	Thesis Structure	35
2	Retrieving RNA gene alignments from genomic data	37
2.1	Introduction	37
2.1.1	Sources of RNA data	37
2.1.2	Genomic alignments	39
2.1.2.1	MultiZ genomic alignments	39
2.1.2.2	EPO genomic alignments	40
2.1.2.3	Evaluating genomic alignments of RNA genes	40
2.1.3	The MARMOSSET pipeline	41
2.2	Materials and methods	42
2.2.1	MonkeyShines software library	42
2.2.1.1	MonkeyShines Perl modules and example functions	42
2.2.1.2	Visualising alignments with BABOON	43
2.2.2	MARMOSSET pipeline structure	45
2.2.2.1	Rfam data	48
2.2.2.2	Find genomic locations in reference	49
2.2.2.3	Fetch genomic alignments	49
2.2.2.4	Examine sequence context	50
2.2.2.5	Perform structural re-alignments	51
2.2.2.6	Generate alignment statistics	51
2.2.2.7	Create neighbour-joining trees	51
2.2.3	Source code	52
2.3	Results and discussion	52
2.3.1	MARMOSSET pipeline execution	52
2.3.1.1	Rfam data	52
2.3.1.2	Find genomic locations in reference	52
2.3.1.3	Fetch genomic alignments	53
2.3.1.4	Examine sequence context	57
2.3.1.5	Generate alignment statistics	57

2.3.2	RNA gene properties	58
2.3.3	Comparing alignments	60
2.3.3.1	EPO and MultiZ alignments	60
2.3.3.2	Structural re-alignments	61
2.3.4	Conclusions	62
3	Evaluating <i>de novo</i> RNA gene prediction software	64
3.1	Introduction	64
3.1.1	<i>De novo</i> RNA gene prediction software	64
3.1.1.1	RNAz	65
3.1.1.2	EvoFold	67
3.1.2	Genome scans	68
3.1.3	Evaluating RNA gene prediction	68
3.2	Materials and methods	70
3.2.1	Data	70
3.2.1.1	Genomic alignments of RNA genes	70
3.2.1.2	Randomised alignments	70
3.2.1.3	Phylogenetic trees	71
3.2.2	RNA gene prediction programs	71
3.2.2.1	EvoFold	71
3.2.2.2	RNAz	72
3.2.3	Evaluation	72
3.2.3.1	Predicting RNA genes	72
3.2.3.2	Detecting RNA gene boundaries	73
3.2.4	Implementation	73
3.3	Results	74
3.3.1	True and false positive rates of RNA gene prediction	74
3.3.2	Detecting RNA gene boundaries	74
3.3.3	Comparing prediction programs	76
3.3.4	Properties that affect RNA gene prediction	77
3.4	Discussion	77
4	Comparing evolutionary models across state space	81
4.1	Introduction	81
4.2	Comparing evolutionary models across state space	82
4.2.1	Modelling evolution as a Markov process	82

4.2.2	Mapping between compound and distinct models	83
4.2.2.1	Definitions	83
4.2.3	Distinct models to compound models	84
4.2.4	Compound models to distinct models	86
4.2.5	Equivalent compound and distinct models	91
5	Investigating RNA models of evolution	94
5.1	Introduction	94
5.2	Materials and methods	97
5.2.1	Substitution models	97
5.2.1.1	Definitions	97
5.2.1.2	Nucleotide and dinucleotide models	97
5.2.1.3	Modelling RNA evolution	100
5.2.2	Model comparison	101
5.2.2.1	Comparing 4-state and 16-state models	102
5.2.2.2	Comparing 7-state and 16-state models	102
5.2.3	Implementation	104
5.2.3.1	Modifications to the PHASE software	104
5.2.4	Genomic alignments of RNA genes	105
5.3	Results	105
5.3.1	RNA models describe evolution better than DNA models	105
5.3.2	Factors determining model choice	108
5.3.3	Model choice affects tree inference	109
5.4	Discussion	111
6	Conclusions	115
	References	119
A	Mapping eigenvalues from compound to distinct models	142
B	Example of calculations for the RY model	145
C	Dinucleotide model definitions	148
D	Modifications to the PHASE software	165

Word Count: 29,843

List of Tables

2.1	Nucleotide colour scheme.	45
2.2	Results of filtering human genomic locations.	53
2.3	Species in EPO alignments (Ensembl release 67).	54
2.4	Results of filtering genomic alignments.	55
2.5	RNA types in the filtered datasets.	59
3.1	Proportion of identical bases in genomic alignments and randomisations.	71
3.2	Predicting RNA genes.	73
3.3	Statistical definitions.	73
3.4	Detecting RNA gene boundaries.	73
3.5	Mean sensitivity and specificity in detecting RNA gene boundaries.	75
5.1	Number of models with $\Delta\text{AICc} = 0$	106
5.2	Best-fit models for EPO-35 alignments, classified by RNA type.	108
C.1	Model 7A	149
C.2	Model 7B	150
C.3	Model 7C	151
C.4	Model 7D	152
C.5	Model 7E	153
C.6	Model 7F	154
C.7	Model 7G	155
C.8	Model 16A	156
C.9	Model 16B	157
C.10	Model 16C	158
C.11	Model 16D	159
C.12	Model 16E	160
C.13	Model 16F	161

C.14 Model 16I	162
C.15 Model 16J	163
C.16 Model 16K	164
D.1 Options for calculating frequencies of base pairs in which one member is a gap.	167

List of Figures

1.1	An example of a microRNA.	20
2.1	A four-species alignment of a transmembrane protein.	46
2.2	A ten-species alignment of an intronic snoRNA and flanking sequence.	47
2.3	MARMOSSET pipeline structure.	48
2.4	Examples of EPO-12 alignments of RNA genes with 400 flanking bases.	56
2.5	Gap length distributions for RNA gene regions and flanking regions.	61
2.6	Pairwise difference between species in EPO and MultiZ alignments.	62
2.7	Distance between EPO genomic alignments and structural re-alignments.	63
3.1	True and false positive rates of RNA gene prediction.	75
3.2	Detecting RNA gene boundaries in EPO-35 alignments.	76
3.3	Overlap of predictions between EvoFold and RNAz.	77
3.4	The effect of GC content, paired bases, and species number on RNA gene prediction.	78
4.1	A four taxa, distinct-state, tree.	92
5.1	Summary of the parameters and relationships of RNA and DNA models.	99
5.2	Graphical representations of example substitution models.	100
5.3	Schematic of mapping 4-state and 7-state models to 16-state space.	101
5.4	Distribution of AICc values relative to the best fit model.	107
5.5	Factors affecting model choice.	109
5.6	Effect of model choice on tree inference.	111

Abstract

USING PHYLOGENETICS AND MODEL SELECTION TO INVESTIGATE THE EVOLUTION OF RNA GENES IN GENOMIC ALIGNMENTS

James Edward Allen

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy, 2013

The diversity and range of the biological functions of non-coding RNA molecules (ncRNA) have only recently been realised, and phylogenetic analysis of the RNA genes that define these molecules can provide important insights into the evolutionary pressures acting on RNA genes, and can lead to a better understanding of the structure and function of ncRNA. An appropriate dataset is fundamental to any evolutionary analysis, and because existing RNA alignments are unsuitable, I describe a software pipeline to derive RNA gene datasets from genomic alignments. RNA gene prediction software has not previously been evaluated on such sets of known RNA genes, and I find that two popular methods fail to predict the genes in approximately half of the alignments. In addition, high numbers of predictions are made in flanking regions that lack RNA genes, and these results provide motivation for subsequent phylogenetic analyses, because a better understanding of RNA gene evolution should lead to improved methods of prediction.

I analyse the RNA gene alignments with a range of evolutionary models of substitution and examine which models best describe the changes evident in the alignment. The best models are expected to provide more accurate trees, and their properties can also shed light on the evolutionary processes that occur in RNA genes. Comparing DNA and RNA substitution models is non-trivial however, because they describe changes between two different types of state, so I present a proof that allows models with different state spaces to be compared in a statistically valid manner. I find that a large proportion of RNA genes are well described by a single RNA model that includes parameters describing both nucleotides and RNA structure, highlighting the multiple levels of constraint that act on the genes. The choice of model affects the inference of a phylogenetic tree, suggesting that model selection, with RNA models, should be standard practice for analysis of RNA genes.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

- i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.
- ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.
- iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.
- iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see <http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487>), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see <http://www.manchester.ac.uk/library/aboutus/regulations>) and in The University’s policy on presentation of Theses.

Acknowledgements

I would like to thank my supervisor Dr. Simon Whelan for advice and guidance, and CASE supervisor Dr. Nick Goldman, and his group, for an enjoyable and useful visit to the EBI. Thanks also to my advisor Dr. Simon Lovell, past and present members of the Whelan group, and residents of the Bioinformatics corridor in the Michael Smith building, for enlightening discussions, scientific and otherwise. My PhD was funded by the Natural Environment Research Council, with additional CASE funding from the EBI. On a more personal note, my parents have been a constant source of support and encouragement throughout all of my academic endeavours, for which I am deeply indebted and grateful. Thanks are also due to my wife Anna, for her love and understanding, and for moral and practical support, including more than her fair share of child-raising duties, which I will be only too glad to resume. Finally, I would like to thank my son, Toby, for providing a charming reminder of the everyday miracle of evolution, and for gladdening my spirits when I was labouring in a sea of RNA gene alignments.

Chapter 1

Introduction

“Nothing in biology makes sense except in the light of evolution.”

Dobzhansky (1973)

“Much in evolution makes even more sense in the light of phylogeny.”

Avise (2006)

The title of Theodosius Dobzhansky’s 1973 paper is an obvious introduction to a thesis which explores molecular evolution; and the quote from John C. Avise is an apt accompaniment, as this thesis uses a phylogenetic approach for that exploration. The idea of a ‘tree of life’ has been independently proposed many times, in various guises, but the modern phylogenetic interpretation of the phrase stems from the work of Charles Darwin: the only illustration in *On the Origin of Species* is a hypothetical phylogenetic tree (Darwin, 1859). The term ‘phylogeny’, to describe the evolutionary development and relatedness of taxa, was not used by Darwin, but was invented by Ernst Haeckel, who further popularised the idea of evolutionary trees, although his definition of phylogeny was somewhat different to the current meaning (Dayrat, 2003).

Molecular phylogenies are important for the elucidation of the tree of life, and have also been used to address a wide range of biological questions. An interesting and diverse sample of these applications is provided by Avise (2006), including the geographical distribution of chameleons (Raxworthy *et al.*, 2002), the migratory behaviour of thrushes (Outlaw *et al.*, 2003), and the convergent evolution of antifreeze proteins in Arctic and Antarctic fish (Bargelloni *et al.*, 1994). Most of the time the only data available for phylogenetic analyses is from extant species, and from these the complex biological processes of millions (or billions) of years of evolution must

be inferred. Considered in those terms, the progress in the field of molecular phylogenetics in recent decades is remarkable, but there remain a wide range of problems that are yet to be addressed, involving the modelling of a wider variety of biological phenomena and the incorporation of new biological knowledge.

This thesis is concerned with non-coding RNA molecules, whose diversity, in terms of structure and biological function, has only recently been realised. Phylogenetic analysis of the RNA genes that define these molecules requires a different approach compared to protein-coding genes, due to the evolutionary constraints that arise from the conservation of RNA secondary structure. In the following sections of this introductory chapter, I briefly review RNA genes, the secondary structure of non-coding RNA, and the prediction of RNA genes in genomic sequences. A comparative approach is useful in RNA gene prediction, so the following section outlines sequence alignment, with particular respect to its relationship to phylogenetic inference, which is the subject of the subsequent section. The final part of the introduction summarises the aims and objectives of the thesis, which comprises the collation of an appropriate dataset of RNA genes in Chapter 2; an evaluation of the *de novo* prediction of RNA genes in Chapter 3; a method to compare models of RNA gene evolution in Chapter 4; and an application of that method of model comparison in Chapter 5.

1.1 RNA genes

Until recently RNA was viewed primarily as an intermediary molecule in protein production, albeit one which might have originally been responsible for the first life on Earth (Gilbert, 1986; Poole *et al.*, 1998). The function and structure of mRNA, tRNA, and rRNA are well established, but as the volume of genomic data has increased it has become apparent that other forms of RNA are crucial to a wide range of biological processes (Mattick and Makunin, 2006). To distinguish these new RNA from mRNA they were termed non-coding RNA (ncRNA), because they are not directly involved in protein coding. The term is slightly misleading, as RNA genes do code for a biologically functional molecule, just not a protein, but suggestions for an alternative nomenclature (Brosius and Tiedge, 2004) have not been widely accepted. An RNA gene is defined as the DNA sequence which gives rise to an ncRNA molecule, analogous to a protein-coding gene. There is evidence for RNA genes with introns (Meyer, 2007), particularly with regard to the ever-growing repertoire of long non-coding RNA (e.g. Guttman *et al.*, 2009), but here I assume that a single contiguous stretch of DNA

codes for a structured ncRNA molecule.

Some families of ncRNA, such as microRNA precursors, have been relatively well characterised in terms of structure and function (Kozomara and Griffiths-Jones, 2011). However, there are many other poorly understood types of ncRNA that are involved in a heterogeneous array of complex biological processes (Taft *et al.*, 2007; Mercer *et al.*, 2009). RNA genes often have low sequence conservation, which in protein-coding genes may imply a lack of function, but the mechanisms that underlie RNA evolution might be quite different (Pang *et al.*, 2006). In particular, the conservation of RNA secondary structure may be as important, or more so, than sequence conservation, as suggested by the wide range of tRNA sequences which nonetheless have very similar shapes. The act of transcription itself (rather than a biologically functional product) has been proposed as an explanation for the lack of sequence conservation, but this seems plausible only for a small part of the transcription activity (Mercer *et al.*, 2009). Nonetheless, RNA secondary structure has not been thoroughly examined with a phylogenetic approach for types of ncRNA other than rRNA and tRNA, and extrapolation from analyses of these molecules may be inappropriate because their fundamental biological roles make them atypical examples of RNA genes.

1.1.1 Types of non-coding RNA

The range of non-coding RNA is increasingly large and diverse, and not all types are pertinent to this thesis. For example, since the methods applied in this thesis explicitly aim to exploit information about secondary structure, in this section I review some important types of structured RNA, but omit details of unstructured RNA such as *Xist* (Brown *et al.*, 1992) or piRNA (Girard *et al.*, 2006; Watanabe *et al.*, 2006). Also, because I use vertebrate datasets, RNA types that are specific to other taxonomic groups (e.g. bacteria: Waters and Storz, 2009, or plants: Schwach *et al.*, 2009) are not covered.

Ribosomal and transfer RNA have central roles in protein synthesis and have been widely studied, but a large repertoire of ncRNA is now acknowledged, and Eddy (2001) has written a comprehensive historical review of the field. The review by Mattick and Makunin (2006) also provides a good overview of the diversity of RNA genes. But, any discussion of ncRNA must start with a brief review of rRNA and tRNA.

Ribosomal RNA are part of the ribosome, the ribonucleoprotein complex responsible for protein synthesis. The ribosome consists of small and large subunits (SSU and LSU, respectively); the SSU binds to mRNA, and the LSU binds to tRNA and generates a polypeptide. In eukaryotes there is one rRNA in the SSU and three in the LSU.

The rRNA molecules provide the core structure of the ribosome, contain binding sites for tRNA (Yusupov *et al.*, 2001), and catalyze the formation of peptide bonds (Noller, 2005). Transfer RNA are quite small (up to 100 bases) and have a highly conserved clover leaf structure. The stem of a tRNA binds to an amino acid that corresponds to the anticodon sequence of the loop, which the ribosome will match to an mRNA codon. The biogenesis of tRNA is dependent on another ubiquitous and ancient RNA type; RNase P RNA is part of a ribonucleoprotein responsible for the cleavage of sequence from the 5' end of primary tRNA transcripts (Guerrier-Takada *et al.*, 1983; Evans *et al.*, 2006).

Ribosomal RNA biogenesis occurs in the nucleolus, and the necessary chemical modifications are guided by members of large family of small nucleolar RNA (snoRNA) genes (reviewed in Granneman and Baserga, 2004), which may also act on some tRNAs (Kiss, 2001). There are two main types of snoRNA, each of which acts as part of a ribonucleoprotein (snoRP): C/D-box snoRPs perform methylation, and H/ACA-box snoRPs perform pseudouridylation. There is also evidence that snoRNA have roles beyond guiding chemical modifications, such as alternative splicing (Kishore and Stamm, 2006).

It is clear that RNA molecules often work in concert with proteins, and other examples of ribonucleoproteins include the relatively well-known structure of the spliceosome, and the more mysterious Vault complex. Like the ribosome, the major and minor spliceosomes are large, biologically fundamental complexes of proteins and RNAs; each spliceosome contains five spliceosomal RNAs. The spliceosomal RNAs recognise the boundaries between introns and exons, and also have a catalytic role (Valadkhan, 2005). In the same way that snoRNAs guide chemical modifications of rRNA, small Cajal body RNAs (scaRNAs) guide methylation and pseudouridylation of spliceosomal RNA, which occurs in a nuclear organelle called the Cajal body (Darzacq *et al.*, 2002). The Vault organelle is a ribonucleoprotein that is conserved in eukaryotes but whose function is only partially understood (Kedersha and Rome, 1986). Vault RNA produces small RNA molecules (in much the same way that precursor miRNAs produce miRNA, albeit through a different pathway), and the effect on gene expression of these RNA molecules may be the cause of the drug resistance associated with the Vault complex (Persson *et al.*, 2009).

Although many types of non-coding RNA are involved in the function of ribonucleoproteins, other types, notably microRNA (miRNA), have a role in the regulation

and modification of protein-coding genes. MicroRNAs are short sequences that down-regulate gene expression through a range of mechanisms (Morozova *et al.*, 2012), and are derived from the cleavage of a precursor (pre-miRNA), which has a hairpin structure. The pre-miRNAs are themselves derived from a higher-level structure, a primary transcript (pri-miRNA), which often exists within the introns of protein-coding genes. The history and biogenesis of miRNA are reviewed in He and Hannon (2004). MicroRNA are now widely studied due to their prevalence (Friedman *et al.*, 2009) and role in disease (e.g. Valastyan *et al.*, 2009; Trajkovski *et al.*, 2011), and also because they are often highly conserved and are thus informative in evolutionary analyses (e.g. Sperling *et al.*, 2009; Heimberg *et al.*, 2010; Shen *et al.*, 2011).

In this thesis I use the term ‘RNA gene’ in a broad sense that includes cis-acting RNA structures in UTRs, which, like miRNA, often have regulatory functions. For example, selenocysteine insert sequences (SECIS elements) have a characteristic structure that causes UGA stop codons to be translated as selenocysteines (Walczak *et al.*, 1996); and iron-response elements (IRE) bind to certain proteins to control the translation of iron-dependent genes, based on the concentration of iron in the cell (Address *et al.*, 1997).

Many of the RNA types described in this section so far have quite well-defined functions, and many are also relatively short. In contrast, long non-coding RNA (lncRNA) are defined (arbitrarily) by a length greater than 200 bases, rather than functional or structural similarity, and thus represent a heterogeneous, but possibly very large (Carninci *et al.*, 2005), group of RNA. In general, the functions of lncRNA are less well-understood than those of shorter types such as snoRNA and miRNA, but examples include chromatin modification, and the transport and activation of transcription factors (reviewed in Mercer *et al.*, 2009; Wang and Chang, 2011; Rinn and Chang, 2012). Long ncRNA genes may have alternating structured and unstructured regions, giving an exon/intron-like pattern. So, although the whole sequence may not be amenable to the analyses in this thesis, the exon-like regions can be treated in the same manner as the shorter RNA genes, under the assumption that the secondary structure of the RNA underlies its function, and thus its evolution.

1.1.2 RNA secondary structure

Although some ncRNA is unstructured, most well-known ncRNA have structures that are primarily composed of short helices of paired nucleotides (‘stems’) connected by

single-stranded regions ('loops'). Hydrogen bonds are formed between complementary base pairs in stems to give the structure stability, which enhances the stability arising from stacking interactions between neighbouring base pairs. 'Watson-Crick' base pairs in RNA structures are defined as A-U, U-A, C-G and G-C pairs; G-U and U-G pairs are known as 'wobble' base pairs, due to their slightly less stable nature; and together, these pairs are 'canonical' base pairs.

Conservation at the structural level in RNA genes occurs through the maintenance of helices in which a pair (rather than either of the paired nucleotides) is conserved by compensatory mutations. For example, a G-C base pair in an RNA gene could mutate from G-C to G-U, which is slightly less stable, but perhaps not sufficiently so to be removed by purifying selection; a subsequent mutation might then create an A-U pair (or revert back to G-C).

Non-coding RNA can fold into more complex 3-D structures, but compared to RNA secondary structure, tertiary structure has not been extensively studied (Holbrook, 2008). This is beginning to change, as the volume of experimentally-determined structures increases in tandem with new computational methods for prediction (e.g. Rother *et al.*, 2011; Bida and Maher, 2012), but I focus here on secondary structure because it clearly underlies the tertiary structure, and the evolutionary conservation of base pairs suggests that secondary structure is a dominant force in RNA evolution. Figure 1.1 shows an example of the secondary structure of a microRNA.

The prediction of RNA secondary structure from primary sequence has been an active area of research for more than 25 years (and the same concepts and methodology also underlie the prediction of RNA genes, of which more later). From an evolutionary perspective, the most interesting approaches to secondary structure prediction are 'phylo-grammars' that explicitly use a phylogenetic tree and probabilistic models of evolution (Knudsen and Hein, 2003; Klosterman *et al.*, 2006; Barquist and Holmes, 2008). The concept of formal grammars, from the field of linguistics (Chomsky, 1959), has been applied to RNA structure prediction because the base pairing behaviour of RNA is well modelled by stochastic context-free grammars (SCFGs) (Sakakibara *et al.*, 1994; Durbin *et al.*, 1998). These approaches can perform with comparable efficiency to popular methods based on experimentally-derived thermodynamic parameters (Dowell and Eddy, 2004), and can, in fact, incorporate such parameters (Rivas *et al.*, 2012). (Note, however, that SCFGs cannot accommodate pseudoknots in secondary structures.) Having a better understanding of RNA gene evolution could lead to more biologically realistic phylo-grammars that have the potential to improve

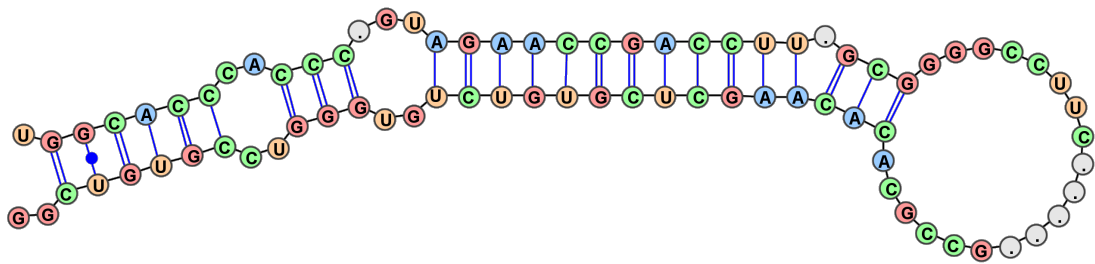


Figure 1.1: An example of a microRNA. The consensus secondary structure shown here is from the Rfam (version 10.1: Gardner *et al.*, 2011) seed alignment of the mir-10 microRNA (RF00104), with sequence data from *Homo sapiens* (EMBL: AC018755.3). Hydrogen bonds between base pairs are shown as bars connecting bases; a wobble base pair is indicated with a dot on a single bar. The sequence data includes gaps, shown here as grey circles, and indicates variability in the length of the loop region in different species. The effect of deriving the structure from multiple sequences is also apparent in the second bulge loop, where there are apparently two, unconnected, wobble base pairs; in fact, one side of the loop is highly variable, and non-complementary pairings in other species mean that the bases are unpaired in the consensus structure. This raises the interesting question, which could be answered with the aid of phylogenetic information, of whether other, non-human, primates have complementary base pairs in this region, which would suggest relatively recent coevolution in that group, towards a structure with a longer main helix. The image was created with VARNA (Darty *et al.*, 2009).

the accuracy of both secondary structure prediction and RNA gene prediction (Dowell and Eddy, 2006; Anderson *et al.*, 2012).

1.1.3 RNA gene prediction

There may exist unknown families of non-coding RNA, and unidentified members of existing families, and their discovery has the potential to open up important new areas of research, as was the case with the role of microRNAs in cancer (Oulas *et al.*, 2009), for example. Predicting the location of RNA genes from genomic sequence data, however, is difficult. RNA genes are heterogeneous and lack the well-characterised signals of protein-coding genes, such as open reading frames. The genetic signatures of ncRNA, such as conservation of promoters or altered expression patterns in disease (Mattick, 2009), are an active area of research, but the discovery of new types of ncRNA is restricted by a lack of in-depth knowledge about what to look for.

There are two main methods of finding new RNA genes, through homology with known data, or by *de novo* approaches that rely on RNA secondary structure prediction and comparative analysis. Meyer (2007) reviews and compares these two approaches, and emphasizes the need to develop a more complicated and nuanced understanding of the concept of an RNA gene when attempting to predict new types of ncRNA.

Finding homologs of RNA genes is non-trivial and computationally expensive, since the comparison must extend beyond sequence similarity to structural similarity. Menzel *et al.* (2009) highlight the difficulties involved in detecting RNA gene homologs, and conclude that without structural alignments, simply using BLAST (Altschul *et al.*, 1997) may be the best method for most biologists. In this thesis I study *de novo* RNA gene prediction (reviewed by Gorodkin *et al.*, 2010) because it can be less computationally-intensive than homology searching, and has the potential to find novel ncRNA families and divergent members of existing families.

The many algorithms and methods that have been developed to predict secondary structure from aligned sequences suggest methods for finding RNA genes (Bernhart and Hofacker, 2009; Gorodkin and Hofacker, 2011). There are three prominent methodologies for secondary structure prediction: programs either (i) use thermodynamic information, (ii) predict structure and alignment simultaneously, or (iii) use probabilistic and evolutionary models. Underlying the last two of these methods is the assumption that coevolutionary processes maintain structure rather than sequence, predominantly through correlated mutations in the helical stems of the RNA that conserve canonical base pairs rather than individual bases. All of these techniques are typically

applied to multiple sequence alignments, although they may differ in the way that they exploit the additional information available in a comparative analysis.

Simultaneous structure and alignment prediction is computationally expensive and is currently not practical for scanning multiple sequence alignments for RNA genes, although good results have been reported for pairwise alignments by using heuristics to reduce computation time (Havgaard *et al.*, 2007; Uzilov *et al.*, 2006). Thermodynamic approaches are based around the minimization of the free energy of an RNA molecule across all of the sequences in an alignment, using experimental data to estimate the energies involved. To generate predictions, the free energy results are combined with information on conservation of sequence and structure (Coventry *et al.*, 2004; Washietl *et al.*, 2005b; Gruber *et al.*, 2010). Probabilistic models have been central to searching for homologs of known RNA genes (Klein and Eddy, 2003; Nawrocki *et al.*, 2009; Gardner, 2009), but have also been applied to *de novo* searches (Rivas and Eddy, 2001; Pedersen *et al.*, 2006; Yao *et al.*, 2006; Bradley *et al.*, 2009b). The probabilistic methods use SCFGs to model secondary structure and sequence evolution, and may also require an *a priori* phylogenetic tree. In Chapter 2, I evaluate thermodynamic and probabilistic programs for RNA gene prediction, both of which require an existing multiple sequence alignment.

1.2 Sequence alignment

Historically, the starting point of many bioinformatic analyses, including phylogenetics, was a fixed alignment of some homologous sequences. Generating such alignments is a complex problem, and will not be extensively addressed here (see Durbin *et al.* (1998) for a historical review; Kemena and Notredame (2009) for a more recent review; Felsenstein (2004, pp.496-520) for a phylogenetics-focused summary; and Morrison (2009b) for a critique of the current usage of alignment in phylogenetics). Instead, I will briefly review alignment in a phylogenetic context, before discussing the genomic alignments that are the basis of the RNA gene datasets that I use in subsequent chapters.

1.2.1 Multiple sequence alignment and phylogenetics

In a phylogenetic context, a multiple sequence alignment is an arrangement of sequences that reflects evolutionary events, rather than, say, structural similarity (Morrison, 2009a). Thus, it is usually necessary to distinguish between orthology and paralogy, as the different types of homology have different phylogenetic interpretations; but the parameters of alignment software are typically tuned with reference to structural alignments (Kemena and Notredame, 2009), which do not necessarily reflect the evolutionary events that interest phylogeneticists.

Treating alignment and phylogenetic inference as two distinct steps allows phylogenetic analysis of existing alignments, and until recently was the only computationally feasible approach. However, alignment and phylogenetic inference are not independent, and both can be improved by calculating phylogenetically-informed alignments (Löytynoja and Goldman, 2005, 2008). An alignment and a tree can be co-estimated in a probabilistic framework (Thorne *et al.*, 1991, 1992; Suchard and Redelings, 2006; Novák *et al.*, 2008) or with heuristic methods (Bradley *et al.*, 2009a), potentially improving the accuracy of both, and providing a measure of alignment uncertainty analogous to the standard practice of evaluating the reliability of tree inference. These sophisticated alignment methods are tractable but complex and computationally demanding, so the future of sequence alignment might involve the use of an initial approximate method, whose results are refined with a more biologically realistic approach (Paten *et al.*, 2009). Whatever method is used, it is important to acknowledge the correctness of the sequence alignment as an assumption in subsequent phylogenetic inference (Löytynoja and Goldman, 2008; Wong *et al.*, 2008).

1.2.2 Genomic alignment

Genomic alignments are multiple sequence alignments of genomes (in practice, ordered sets of alignments of genomic regions), and can be useful for analysis of regions that do not contain protein-coding genes. Multiple sequence alignments of protein-coding sequences typically consider a limited set of small-scale mutations, such as single-base changes, insertions, and deletions. In order to align genomes, larger-scale evolutionary events, such as duplications and inversions, must be modelled. Genomic alignment methods usually make the simplifying assumption that the small-scale and large-scale evolutionary events are independent, and use multi-step pipelines to first cluster homologous regions and then refine the alignment (Kent *et al.*, 2003).

Compared to the range of alignment software for genes, there are relatively few programs that produce genomic alignments, and there are only two that have associated datasets available for public consumption, MultiZ and EPO. MultiZ is part of the TBA (Threaded Blockset Aligner: Blanchette *et al.*, 2004) pipeline, and combines pairwise alignments generated by BlastZ (Schwartz *et al.*, 2003) or LastZ (Harris, 2007) to generate a multiple alignment. A MultiZ genomic alignment is a connected set of local alignments, ('blocks') arranged (or 'threaded') with respect to a reference species. It is possible to use TBA to extend these MultiZ alignments to a non-reference-based alignment, but this is not practicable for whole genomes.

The EPO (Enredo-Pecan-Ortheus) pipeline has three stages, the first of which (Paten *et al.*, 2008a) creates an iteratively-refined graph that clusters homologous regions. These regions are then aligned by Pecan (Paten *et al.*, 2008a, 2009), which is an extension of the consistency-based optimization function in the ProbCons (Do *et al.*, 2005) alignment software. Finally, the Ortheus program (Paten *et al.*, 2008b) is used to infer ancestral genome sequences.

More recent methods of genomic alignment use a more nuanced hierarchical representation that reflects evolutionary events as a continuum, rather than a false, albeit useful, dichotomy between small-scale and large-scale events (Paten *et al.*, 2011a). These have the potential to provide more accurate alignments than MultiZ or EPO (Paten *et al.*, 2011b), but genomic alignment datasets have not yet been made available via the UCSC Genome Browser.

1.3 Phylogenetic inference

Phylogenetic inference is used to generate a tree, and possibly the parameters of an evolutionary substitution model, from a sequence alignment, and inference techniques fall into three broad categories. Distance matrix methods are the simplest and make the least efficient use of the data (in that the difference between two sequences is reduced to a single value), but remain useful for approximations and initial analyses. Parsimony was developed at around the same time, and was once *de rigueur*, but has fallen out of favour somewhat. In recent years the parametric techniques of maximum likelihood and Bayesian inference have become the most prominent.

1.3.1 Distance matrices

Distance matrix methods calculate pairwise distances between taxa in a sequence alignment, then aim to construct a tree for all of the taxa that best approximates those distances. There are a variety of methods (Desper and Gascuel, 2007), of which UPGMA (Sokal and Sneath, 1963) and neighbour-joining (Saitou and Nei, 1987; Gascuel, 1997) are the most widely used. Neighbour-joining remains popular as a quick means of generating a reasonable tree, which can be used as a starting point for more sophisticated methods. Significant problems with all distance methods are that they discard phylogenetic information, and use information about rate variation significantly less efficiently than likelihood methods (Felsenstein, 2004).

1.3.2 Parsimony

Parsimony methods are character-based, rather than distance-based; it is quite difficult to formally define exactly what a ‘character’ is in general (DeSalle, 2006), but in a molecular context a character is a base or an amino acid (or a gap) in a sequence alignment. Parsimony methods seek the tree in which the least evolution has occurred, meaning the one with fewest changes of character (Fitch, 1971), and this intuitively attractive aim, along with computationally tractable algorithms, saw the method gain popularity throughout the 1980s and 90s. The main problem with parsimony is that it can be inconsistent under certain conditions; that is, as more data is added, the method becomes increasingly certain that a wrong tree is correct (Felsenstein, 1978). This behaviour is sometimes termed ‘long branch attraction’, although that phrase can be misleading, as long branches are not a necessary or sufficient characteristic of inconsistency (Kim, 1996). Other problems with parsimony are that it is generally used in a non-parametric manner (so it is not amenable to robust statistical analysis), and that it does not involve an explicit model of evolution.

1.3.3 Maximum likelihood and Bayesian analysis

Maximum likelihood (ML) and Bayesian analysis are both parametric techniques which share many characteristics, particularly when contrasted to parsimony. Thorne *et al.* (1992, p.4) eloquently justify the use of these methods (with reference to ML, but also applicable to Bayesian techniques): “The advantages of this approach include explicit assumptions, a model of sequence change based upon actual biological phenomena

instead of arbitrary criteria for sequence comparison, and the vast statistical theory concerned with likelihood methods”.

Working within the rigorous mathematical framework of statistical inference is appealing because statistical hypothesis testing can provide a degree of confidence in the results (Goldman, 1993). The major downside of parametric techniques is the level of computation required, even for moderate numbers of taxa. Finding the best tree by exhaustive search is generally not possible, so heuristic methods are used (this may also be the case for parsimony methods). Heuristics are not guaranteed to find the globally optimum tree, but in practice they work quite well (Whelan, 2007).

Maximum likelihood methods became prominent in phylogenetics after Felsenstein (1981) proposed a pruning algorithm (a dynamic programming technique) that allowed for computation to be completed in a reasonable time. ML was not in common use until the mid-1990s, however, probably due to a combination of inertia from researchers who were accustomed to parsimony, and computational speed; the method is tractable, but still requires substantial number-crunching. The use of maximum likelihood methods in phylogenetics are comprehensively reviewed by Felsenstein (2004) and Buschbom and von Haeseler (2005).

The application of Bayesian theory to phylogenetics stems from the mid-1990s (Rannala and Yang, 1996; Yang and Rannala, 1997), but the approach was not widely used until a review for a general audience (Huelsenbeck *et al.*, 2001) and user-friendly software were written (MrBayes: Huelsenbeck and Ronquist, 2001; Ronquist *et al.*, 2012). Bayesian inference uses the same definition of likelihood as ML, but aims to compute the posterior probabilities of the tree and model parameters, rather than a maximum likelihood estimate (Huelsenbeck and Ronquist, 2005). Exact calculations are not generally practicable in a Bayesian approach, but Markov chain Monte Carlo (MCMC) methods (Huelsenbeck *et al.*, 2001) and sequential Monte Carlo methods (Bouchard-Côté *et al.*, 2012) are effective in producing good approximations.

1.3.4 Choosing an inference method

Each inference method has a band of advocates who favour methods on ideological or philosophical grounds, which is often difficult to disentangle; for example, Karl Popper has been invoked on both sides of the parsimony-versus-likelihood debate (Helfenbein and DeSalle, 2005; Faith, 2006). In practice, it is perhaps best to take a pragmatic approach and focus on the properties of the different methods, and the situations in which they may outperform one another in practice.

Distance-matrix methods are good for quick approximations, but are too simplistic for thorough phylogenetic inference. Parsimony can become inconsistent, but if the number of changes per site is small (that is, the rate of evolution is slow) then the method can perform well, and approximates maximum likelihood (Steel and Penny, 2000). Parsimony and ML can, in fact, be conceptualised as the extreme ends of a sliding scale (Tuffley and Steel, 1997), and it has been shown (for a small tree, as the computations are complex) that regions of the parameter space in which one method is inconsistent are consistent for the other method (Kim and Sanderson, 2008). Parsimony can be useful because ML (and Bayesian) calculations are computationally intensive, which may be prohibitive for large numbers of taxa.

The computational requirements of the parametric methods can be assuaged by various clever heuristics and approximations, but model misspecification can also be a problem. ML is only consistent if the model is adequate, and determining what represents an 'adequate' model is not easy; although the method tends to be fairly robust to model misspecification (Yang *et al.*, 1994; Kelchner and Thomas, 2007). If many parameters are being estimated then the advantages of taking a statistical approach are lessened, since statistical power is inversely proportional to the number of parameters.

The robust, statistical basis of ML and Bayesian methods have made them popular in phylogenetic inference, but choosing between them is rather difficult, perhaps because of their similarities (akin to choosing between a tangerine and a satsuma, compared with the banana of parsimony). The two methods have quite different underlying philosophies, however, and one of the main criticisms of Bayesian methods has been the need to specify prior probabilities which do not always have a clear biological meaning. In practice, this may be less important than the quantifiable phylogenetic uncertainty that the Bayesian approach inherently provides, but which ML lacks (bootstrapping methods can be used, but this requires additional work). Even then, however, there is debate over the accuracy of the posterior probabilities produced by the Bayesian methods, which are used to indicate statistical confidence (Kolaczowski and Thornton, 2007). The parametric methods have the advantage of an explicit evolutionary model, meaning that conclusions can be drawn about the biology of evolution; but this will not always be relevant in a study, in which case the parameters may add unnecessary complexity (Edwards, 1972).

1.4 Substitution models of evolution

Mathematical models of molecular evolution describe the probability that one state (e.g. nucleotide, amino acid) changes to another; a matrix containing these probabilities is known as a transition rate matrix. Each model has some specific assumptions, but other assumptions are more general. Unless otherwise noted, all the models that are described below assume stationarity, homogeneity, and reversibility. Stationarity and homogeneity imply that the nucleotide frequencies and the rate of substitution, respectively, have remained constant throughout evolution. Reversibility indicates that a change in one direction (e.g. from A to C) is as likely as a change in the opposite direction (from C to A). Another general assumption is that each site in a sequence evolves at the same rate, and independently of other sites. Liò and Goldman (1998) and Whelan *et al.* (2001) review the most popular nucleotide and amino acid models in detail, and Delport *et al.* (2009) provide a review of codon-based models.

1.4.1 DNA models

The simplest model for the evolution of DNA is the Jukes-Cantor model (Jukes and Cantor, 1969), usually known more succinctly as the JC or JC69 model. Each base has an equal chance of changing to another, and the bases are assumed to exist at equal frequencies throughout the DNA sequence. The JC model is, in most cases, too simple to model biological data, but its importance lies in being the first evolutionary model to account for unobserved mutations, through the treatment of evolution as a Markov process. This is biologically reasonable; once a base has mutated, subsequent mutations will depend only on the current base, and knowledge of the previous bases is not available to evolutionary processes.

Kimura (1980) generalised the JC model by assigning different rates of change for transitions and transversions (the K80 or K2P model); all bases have the same equilibrium frequencies, as in the JC model. Felsenstein (1981) created a different generalisation of the JC model (the F81 or FEL model), by keeping the rate of change the same for all bases, but allowing unequal expected frequencies. These two approaches were combined in the HKY model (Hasegawa *et al.*, 1985), and this is probably the simplest model that is currently in relatively common usage; the computational requirements are greater than for the JC, K80, or F81 models, but are not onerous for modern technology.

Taking this process of generalisation further gives the general time-reversible (GTR)

model, also referred to as REV. In this model the equilibrium frequencies can be unequal, and six parameters define the rates of change between each permutation of base pairs. Acknowledging the development of the GTR model is not entirely straightforward; Lanave *et al.* (1984) were the first to use such a model, but as Yang (1994a) noted, they stated that the model was general but implicitly assumed reversibility. Perhaps for that reason, Tavaré (1986) is sometimes credited with the invention of the GTR model; this author also provided a transition rate matrix for the model and framed it in the context of previous models, which Lanave *et al.* did not. As a further complication, Yang's lucid exposition of the model (1994a) is occasionally cited instead.

If the restriction on reversibility is removed from the GTR model, then the most general model has 12 independent substitution parameters (Rodríguez *et al.*, 1990), but this model is very difficult to work with (Yang, 1994a; Felsenstein, 2004, p.210-211), and there is scant evidence on the level of improvement over the GTR model (but see the section below, on model assumptions). By constraining different parameters in the GTR model it is possible to generate a suite of evolutionary models. For example, the TN93 model extends the HKY model by considering two types of transition, purine to purine and pyrimidine to pyrimidine (Tamura and Nei, 1993). Zharkikh (1994) provides a review and comparison that includes some of the more esoteric models.

1.4.2 RNA models

Much of the work on DNA models has assumed that sites in a sequence are independent and identically distributed, but this is not the case for RNA, where secondary structure due to complementary base pairs is important. Schöniger and von Haeseler (1994) showed that nucleotides in stem regions are correlated, and that the assumption of no correlation leads to an underestimate of evolutionary distances. A more sophisticated model that treated base pairs in stem regions as the units of evolution, found that these evolved at almost twice the rate of bases in loop regions (Rzhetsky, 1995). Although not explicitly stated in the paper, this implies that pairs of bases are coevolving in order to maintain function.

To account for the dependency introduced by base pairing, RNA-stem models describe changes between pairs of nucleotides (dinucleotides), rather than individual nucleotides (Tillier and Collins, 1995). Savill *et al.* (2001) provided the definitive review of RNA-stem models, grouping them into broad categories based on the states in the model: all dinucleotide combinations (16-state); Watson-Crick and wobble base pairs only (6-state); Watson-Crick and wobble base pairs, with a single 'mismatch' category

representing all other pairings (7-state).

There are several biologically-motivated variations within each of these three groups, such as whether double substitutions (where both nucleotides in a pair change simultaneously) are permitted. In models without double substitutions, compensatory mutations are assumed to proceed explicitly via intermediates that involve single nucleotide changes. There is some evidence that double substitutions occur in nature (Averof *et al.*, 2000; Whelan and Goldman, 2004), but the mechanism of this process is unclear, and it is normally assumed to be an uncommon event (Smith *et al.*, 2003). However, including them in a phylogenetic model can provide a better fit to the data, perhaps because, in some types of RNA at least, compensatory substitutions reach fixation so rapidly they look like a simultaneous substitution in the sequence data (Tillier and Collins, 1998). It may also be that double substitutions are a plausible mechanism at the level of populations rather than individuals (Higgs, 2000).

To analyse an RNA sequence in its entirety, rather than just the stems, it is necessary to partition the sequence into stem and loop regions based on a known structure, and then use a mixture model composed of an RNA-stem model and a DNA model for the loops. An RNA alignment is often associated with a structure, either hand-curated or as ancillary output from RNA-specific alignment software (e.g. Sahraeian and Yoon, 2011), but an alternative approach is to use a mixture model which does not require *a priori* partitioning (Pagel and Meade, 2004; Lanfear *et al.*, 2012).

Software that has been developed for phylogenetic inference with DNA and amino acid models will generally not work with a model of RNA that takes account of base-pairing rules (although MrBayes (Ronquist *et al.*, 2012) does permit a limited set of dinucleotide models). The PHASE program (Jow *et al.*, 2002; Hudelot *et al.*, 2003; Gibson *et al.*, 2005; Gowri-Shankar and Rattray, 2006) has been specifically designed to work with all of the RNA models described by Savill *et al.* (2001), some of which have been incorporated into later versions (7.2.3 and above) of RaxML (Stamatakis, 2006). PHASE is ideally suited to analyses of RNA genes, but has not been actively developed for some time, and in Chapter 5 and Appendix C I describe some modifications and extensions that I made to the program.

1.4.3 Amino acid models

Amino acid models have generally used an empirical approach, generating replacement matrices based on the rates of change from one amino acid to another in groups of related proteins. Dayhoff *et al.* (1978) produced the first empirical model, which

was later updated with more protein data and a faster and less error-prone method, known as the JTT model (Jones *et al.*, 1992). Analyses that use the Dayhoff and JTT models are usually combined with amino acid frequencies estimated from the data being studied, and this is indicated by adding the suffix '+F' to the name of the model.

The counting methods of the Dayhoff and JTT models do not account for multiple substitutions at the same site, and will thus tend to underestimate the overall amount of evolution (the use of closely related proteins in the creation of the replacement matrices is an attempt to alleviate this problem). The WAG model (Whelan and Goldman, 2001) incorporates an approximate method of phylogenetic inference to allow for multiple substitutions, and in most cases it models evolution more accurately than the Dayhoff+F and JTT+F models. Le and Gascuel (2008) incorporated rate heterogeneity across sites (see section 1.4.4) into the WAG model, and also used a larger and more diverse database of protein sequences.

Models for specific types of protein have also been developed to describe, for example, evolution in transmembrane (Jones *et al.*, 1994), mitochondrial (Adachi and Hasegawa, 1996), and retroviral (Dimmic *et al.*, 2002) proteins. A comprehensive study (Keane *et al.*, 2006) demonstrated that no single amino acid model performed the best for all datasets, and that the best model could sometimes be counterintuitive, strongly suggesting the need to use formal methods of model selection (see section 1.4.8).

1.4.4 Rate heterogeneity

The standard models of nucleotide and amino acid evolution assume that all sites in a sequence evolve at that same rate. However, allowing the rate of substitution to differ across the sites in a sequence is much more biologically realistic than assuming a constant rate, as different parts of the sequence may be under very different selective pressures. A successful approach has been to use a gamma distribution to model the rates; in relation to evolutionary models this was first proposed by Uzzell and Corbin (1971), but Yang's computationally tractable treatment (1994b) has become the definitive citation. Yang showed that the fit of the model can improve significantly by considering either a continuous gamma distribution or a set of discrete rate categories drawn from the distribution (see also the review by Yang, 1996).

Along similar lines, it is often useful to consider some sites as invariant (Hasegawa *et al.*, 1985), the assumption being that these sites will be so important to function that any changes will not be viable. These methods can be combined with virtually

any nucleotide or amino acid model, and notationally this is indicated by suffixing the model name with '+I' for invariants, and '+G' or '+ Γ ' for the gamma distribution.

1.4.5 Codon models

The use of codons rather than nucleotides or amino acids in evolutionary models allows for more biological realism, because different positions in a codon evolve at different rates (Bofkin and Goldman, 2007), and because a codon has more information content than an amino acid. The two main types of codon model are named MG (Muse and Gaut, 1994) and GY (Goldman and Yang, 1994), and most subsequent models are a variation or extension of them. Rodrigue *et al.* (2008) review and compare the two methods, and conclude that the MG model performs better than the GY model. Accounting for rate heterogeneity in codon-based models is not straightforward due to the more complicated structure of these models (Anisimova and Kosiol, 2009), and the gamma distribution method is usually not used.

Compared to DNA or amino acid models, the increased complexity of modelling changes between all 61 sense codons is computationally unwieldy, and the methods are not easily applied to phylogenetic inference (Anisimova and Kosiol, 2009). Ren *et al.* (2005) demonstrated that codon models can be applied to moderately-sized datasets, however, and also showed that nucleotide models that allow for codon-position variation (suitable for analysis of larger datasets) can perform well, although this is not necessarily true (Whelan, 2008a).

1.4.6 Context-dependent models

The use of knowledge about codon positions in nucleotide models is a relaxation of the assumption that each site in a sequence is independent of its neighbours. Such context-dependent models have the potential to model evolution more realistically, and although this adds significant complexity, a number of such models have been successful, without finding their way into mainstream use. Hidden Markov Models (HMMs) have been used to model dependence between neighbouring sites with respect to rate heterogeneity (Felsenstein and Churchill, 1996), and protein secondary structure (Goldman *et al.*, 1996). Siepel and Haussler (2004) provide a review of (and rationale for) the application of HMMs to phylogenetics, and extend earlier HMMs to accommodate higher-order states. The known tertiary structure of proteins can also be used to model context dependency (Robinson *et al.*, 2003; Rodrigue *et al.*, 2005),

with interesting results, but the pros and cons of these methods compared to context-independent models are not yet fully explored (Rodrigue *et al.*, 2006). Kosiol *et al.* (2006) show where context-dependency fits in the framework of evolutionary models in general, and note that these approaches are only partially understood.

1.4.7 Model assumptions

The relaxation of the assumptions concerning rate homogeneity and context-independence provide models that perform well, so it is instructive to examine the other common model assumptions. In addition to variation across sites, there may be variation across branches, known as heterotachy (Lopez *et al.*, 2002). The rationale underlying heterotachy is that evolutionary pressures will have varied over time, as environmental conditions have changed, sometimes drastically. In recent years computing power has caught up with theoretical considerations to make heterotachous models viable and informative (e.g. Wang *et al.*, 2007b; Pagel and Meade, 2008; Whelan, 2008b).

The assumptions of reversibility and stationarity are related, and their relaxation is often considered together. Removing these assumptions invalidates many of the tricks used to make computations simpler, so while the most general DNA model was described some time ago (Barry and Hartigan, 1987), it is only recently that practical applications have been developed (Jayaswal *et al.*, 2005; Squartini and Arndt, 2008; Jayaswal *et al.*, 2011). These studies provide evidence that non-reversibility and non-stationarity are a biological reality, and in particular that the GC content may be a significant factor which models should accommodate. However, it is too early to make any general conclusions, and it may be, for example, that GC content is sufficiently correlated with environmental conditions to be modelled by heterotachy. The caveat to all attempts to model evolution in increasingly realistic (and complicated) ways is that a complex model does not necessarily perform better than a simple one, and formal methods of model selection are essential.

1.4.8 Model selection

If an inference method involves an explicit evolutionary model, then a formal model selection procedure should be used to choose an appropriate model (Burnham and Anderson, 2002; Sullivan and Joyce, 2005). Here, ‘appropriate’ means the simplest model (the one with the least parameters) that best fits the data; parameter-rich models will tend to fit the data well, but at the cost of statistical power (Steel, 2005), and

“overparameterization is often the mark of mediocrity” (Box, 1976, p.792).

The likelihood ratio test (LRT) provides a way to compare the maximum likelihood values of two models to determine if one is a statistically significant better description of the data, in a formal hypothesis-testing framework. If the models are nested (that is, one is a general form of the other) then a χ^2 distribution can be used to determine significance (Kendall and Stuart, 1973; Goldman and Whelan, 2000). For non-nested models, and for cases with small sample sizes, the χ^2 approximation is not appropriate, and Monte Carlo simulations are required to calculate significance (Goldman, 1993). LRTs have been applied in a range of phylogenetic scenarios (Huelsenbeck and Rannala, 1997), and can be used in a hierarchical manner to select the best-fitting model (Posada, 2008).

Information-theoretic approaches are an alternative to model selection with hierarchical LRTs (Pol, 2004; Posada and Buckley, 2004). The most popular information-theoretic method uses the Akaike information criterion (AIC) (Akaike, 1974), which includes a term that penalises a model in proportion to the number of its parameters. The Bayesian information criterion (BIC) has a similar formulation, and might provide a stricter test and a clearer interpretation than the AIC, but its suitability requires careful consideration (Posada and Buckley, 2004). Minin *et al.* (2003) suggest a potentially useful extension to the BIC within a decision theory framework, and Sullivan and Joyce (2005) review model selection and place all of the above techniques within such a framework.

1.5 Aims and objectives

The chief motivation for this thesis is the need to develop a better understanding of the evolution of RNA genes, in order to gain knowledge about the biological function of ncRNA molecules. Phylogenetic inference with explicit evolutionary models has been used to investigate the evolution of protein-coding genes, and may also be useful in relation to RNA genes. However, although a variety of different RNA substitution models have been shown to be effective, they are not in widespread use; I suggest that this is, at least in part, because comparisons between DNA and RNA models of evolution are not straightforward. Without a framework to select the model that best fits the data, and subsequently have confidence in its parameterisation and the resultant trees, it is difficult for researchers to justify a move away from the standard practice of applying DNA models. Such model selection is useful not only because the best

models are expected to provide more accurate trees, but also because the properties of the models can shed light on the evolutionary processes that occur in RNA genes. The central aim of this thesis, then, is:

- To present a quick, simple, statistically rigorous way to compare DNA and RNA substitution models of evolution.

Several ancillary objectives accompany that central aim:

- A software pipeline to gather and filter a set of alignments of RNA genes.
- The development of open-source software that enables model selection with DNA and RNA models.
- Analysis of RNA gene alignments to demonstrate the utility of model selection with DNA and RNA models.
- The characterisation of different RNA models in their ability to describe evolution in RNA genes.

Finally, a better understanding of RNA gene evolution should lead to improvements in the *de novo* prediction of RNA genes, and the creation of a large set of RNA gene alignments presents an opportunity for a complementary objective:

- The evaluation of *de novo* RNA gene prediction.

1.5.1 Thesis Structure

Chapter 2 describes the MARMOSSET pipeline that I developed to retrieve, filter, and visualise genomic alignments of RNA genes and their flanking regions. A reusable, flexible pipeline for generating alignments of RNA genes is important in order to keep pace with the rapidly increasing numbers of known RNA genes and genome sequences. The MARMOSSET pipeline is one element of a larger code base, named MonkeyShines, which is also used in subsequent chapters. The MonkeyShines software provides functionality to manipulate and analyse genomic alignments, and includes novel code to create visualisations of alignments with a technique that had been described previously but which had not been implemented elsewhere. In Chapter 3 I test two popular methods of *de novo* RNA gene prediction which have not been evaluated with either large positive datasets or appropriate negative datasets, and assess the factors that can affect performance.

In Chapters 4 and 5 I analyse the RNA gene alignments with a range of evolutionary models, including some specific to RNA evolution, and examine which models best describe the changes evident in the alignment. Chapter 4 consists largely of a mathematical proof that allows models with different state spaces to be compared in a statistically valid manner, extending the work of others who applied the approach to comparisons between nucleotide, amino acid, and codon models. In Chapter 5 I apply the methodology of the previous chapter to a phylogenetic analysis of the RNA gene alignments from Chapter 2, first using maximum likelihood to assess model fit, and then applying Bayesian methods to compare the trees that are generated under different models. To perform these phylogenetic analyses it was necessary to update existing software for phylogenetic inference, to make it work reliably with a range of different types of RNA gene and to implement additional functionality to allow comparison between models with different state spaces. In the final chapter I summarise my results and highlight the advantages of using a phylogenetic approach to investigate the evolution of RNA genes.

Chapter 2

Retrieving RNA gene alignments from genomic data

2.1 Introduction

This chapter describes the methodology that I use to generate alignments of RNA genes, in order to investigate RNA gene prediction and mathematical models of RNA evolution in later chapters. It is necessary to generate alignments because existing alignments of RNA genes never include flanking sequence, which is required to evaluate RNA gene prediction programs because these work by differentiating flanking and gene regions. My RNA gene alignments are the result of a software pipeline that takes the sequence of a known RNA gene, locates it in a reference genome, and then extracts the relevant portion of a genomic alignment. Before describing the pipeline, it is necessary to review the availability of RNA data and the properties of genomic alignments, as these factors affect the interpretation of the RNA gene alignments.

2.1.1 Sources of RNA data

There are a range of publicly available RNA databases, some specialising in particular types of RNA and tending to have manual annotation, while others have a much more general scope and often collate and augment subsets of data from the specialised databases. Specialised databases may contain data based on function (e.g. microRNA in miRBase: Kozomara and Griffiths-Jones, 2011), or structure (e.g. pseudoknots in PseudoBase: van Batenburg *et al.*, 2000; Taufer *et al.*, 2009), and there are comprehensive databases for the study of rRNA (e.g. SILVA: Pruesse *et al.*, 2007). As was

recently noted (Bateman *et al.*, 2011), there is a lack of a comprehensive and canonical source of RNA data, largely due to the relatively recent discovery of the number and variety of RNA genes. Of the general RNA databases, Rfam (Gardner *et al.*, 2011) is widely used, and although it is not the largest such database, its basis in manual curation is useful for my purposes, in which I want to analyse sequences which are almost certainly genuine RNA genes.

Rfam adopts the approach of Pfam (Punta *et al.*, 2012) in using profile-based models to group RNA genes into families. A curated multiple sequence alignment and a consensus base-paired secondary structure is used to produce a covariance model, which defines an RNA family and a set of ‘seed’ alignments. This covariance model can subsequently be used to assign additional RNA genes to that family, to construct the ‘full’ dataset. The database only contains structured RNA, and Rfam alignments do not necessarily represent the evolutionary history of any particular locus.

I use Rfam data as the starting point for my analyses because other general databases focus on a too narrow range of RNA genes (CRW: Cannone *et al.*, 2002), lack structural information (NONCODE: Bu *et al.*, 2012), have been retired (RNAdb: Pang *et al.*, 2007), or add little to the Rfam data (fRNAdb: Mituyama *et al.*, 2009). The BRaliBase datasets (Gardner and Giegerich, 2004; Gardner *et al.*, 2005; Freyhult *et al.*, 2007) are the RNA equivalent of the BaliBase alignments that are widely used to compare sequence alignment programs (Thompson *et al.*, 2005), and are a possible alternative to Rfam as a source of RNA data. However, these datasets only cover a very small number of different RNA genes, and, as with all sources of RNA data, lack flanking sequence. It is important, particularly for RNA gene prediction, to be aware of the genomic context of an RNA gene, and to extend the alignment to flanking regions on either side of the gene. Given this need for flanking sequence, the ENCODE project (ENCODE Project Consortium, 2007; Margulies *et al.*, 2007) is potentially a good source of annotated RNA genes. However, there are only 8 known RNA genes in the pilot ENCODE regions, and 7 of those are either microRNA or one particular type of snoRNA (Washietl *et al.*, 2007). (The results from the most recent ENCODE studies (ENCODE Project Consortium, 2012) were published after the majority of work for this thesis was completed.)

To obtain evolutionary alignments of RNA genes, with flanking sequence, it is thus necessary to generate alignments myself. The basic approach is to find the genomic location of Rfam sequences for a reference species (human in this case) and then retrieve the relevant section of a genomic alignment. Similar methods have been used

before (e.g. Babak *et al.*, 2007), and there is evidence that RNA genes are often well-aligned in genomic alignments (Wang *et al.*, 2007a), but care needs to be taken to appropriately filter and evaluate the resultant RNA gene alignments (Margulies *et al.*, 2007; Chen and Tompa, 2010).

2.1.2 Genomic alignments

The topic of genomic alignment was briefly surveyed in the introductory chapter, and while there several software packages to generate alignments, only two, MultiZ (Blanchette *et al.*, 2004) and EPO (Paten *et al.*, 2008a), have been used to produce large, publicly available datasets, distributed by the UCSC Genome Browser (Dreszer *et al.*, 2012) and Ensembl (Flicek *et al.*, 2012), respectively.

2.1.2.1 MultiZ genomic alignments

The UCSC Genome Browser provides a MultiZ 46-species alignment for vertebrates, with human as the reference species, and this has been used for investigations into RNA genes (e.g. Washietl *et al.*, 2007; Torarinsson *et al.*, 2008; Jeggari *et al.*, 2012) and many other evolutionary analyses (e.g. Washietl *et al.*, 2011; Gelfman *et al.*, 2012; Hiller *et al.*, 2012). There are two potential problems with using MultiZ genomic alignments. The first is that the alignments are now relatively old, given the recent glut of genomic data. The current 46-species vertebrate alignment dates from early 2009, so it lacks certain genomes altogether (e.g. pig, *Sus scrofa*), and many others have been updated (e.g. chimp, gorilla, and cow). This is not an insurmountable problem, as it is possible to generate one's own genomic alignments with the latest genome assemblies, but this is a complicated and time-consuming process.

The second problem with MultiZ alignments is the relatively short block size of the alignments, which means that for most applications multiple blocks need to be joined together. However, if joining the blocks was straightforward and unambiguous, MultiZ would have done it already, so one must decide on criteria that define whether simply concatenating blocks is appropriate or whether more sophisticated processing is required; and these criteria change depending on how the alignment is to be used (e.g. Washietl *et al.*, 2007). Sometimes prior knowledge, such as intron/exon boundaries (Gelfman *et al.*, 2012), can be used to join blocks in a more rigorous manner, but such information is often not available for non-coding regions. Nonetheless, many researchers have joined blocks and successfully used MultiZ alignments, but the problem can

be avoided entirely by using EPO alignments instead.

2.1.2.2 EPO genomic alignments

Ensembl (release 67) provides two mammalian EPO alignments, one with 12 ‘high-coverage’ species, and one with 35 ‘low-coverage’ species which is based on the 12-species alignment; for the sake of brevity I will refer to these as EPO-12 and EPO-35 respectively. The use, by Ensembl, of the terms high-coverage and low-coverage is somewhat misleading, as although the species in the EPO-35 set had 2X coverage when the alignment was first produced, many of the assemblies have subsequently been updated. Moreover, the gorilla assembly in the EPO-12 dataset is officially low-coverage, but is included in the EPO-12 set because it has been assembled onto chromosomes and is considered sufficiently high quality.

EPO alignments have been used to address evolutionary questions (e.g. Romiguier *et al.*, 2010; Scally *et al.*, 2012), but not as extensively as MultiZ alignments, and not, to my knowledge, with regard to RNA genes. For this study, EPO alignments offer a number of advantages over MultiZ alignments. Each release of the Ensembl website (4 or 5 times a year) includes an update to the genomic alignments, enabling the incorporation of new and updated assemblies, in contrast to the less regular updates for MultiZ alignments at the UCSC Genome Browser. The Enredo component of the EPO pipeline generates much longer alignment blocks than MultiZ (Paten *et al.*, 2008a), so blocks need to be joined far less frequently, and alignments that span blocks can be discarded without sacrificing a large amount of data. The Pecan stage of the EPO pipeline has also been shown to produce accurate alignments (Paten *et al.*, 2008a), particularly for non-coding regions (Chen and Tompa, 2010). The EPO alignments cover fewer species than the vertebrate MultiZ alignment, but because alignment quality decreases quite dramatically when moving beyond mammals (Margulies *et al.*, 2007; Chen and Tompa, 2010), it is sensible to restrict this analysis to mammals in any case.

2.1.2.3 Evaluating genomic alignments of RNA genes

The MultiZ and EPO genomic alignments are based on primary sequence, and thus may misalign RNA sequences, in which conservation is often at the structural level. It has been shown that MultiZ produces reasonable alignments of RNA genes (Wang *et al.*, 2007a), and this can be tested for a given data by re-aligning the RNA regions of a genomic alignment with structurally-aware RNA alignment software. There are, naturally, many RNA alignment programs to choose from, and unfortunately there is no

recent review that systematically compares the accuracy of these programs. However, the X-INS-i program in the MAFFT suite of alignment programs (Katoh and Toh, 2008) has been shown to perform well (Sahraeian and Yoon, 2011), and is at least representative of the field. A slight disadvantage of MAFFT/X-INS-i is that it does not output an inferred secondary structure, which would be useful to compare to the consensus Rfam structure. Another RNA alignment program, PicXAA-R (Sahraeian and Yoon, 2011), has similar performance to MAFFT and does provide a structure with each alignment.

In addition to generating alternatives to the genomic alignments, a range of statistics can be calculated for each alignment, and used to judge its quality. These include properties common to all multiple sequence alignments, such as the number of gaps or ambiguous nucleotides, and properties specific to RNA alignments, such as the Structural Conservation Index (SCI: Washietl *et al.*, 2005b). SCI uses minimum free energy calculations to compare the folding energy of the alignment to the average energy of the individual sequences. It is formulated so that values near 0 indicate low conservation, values near 1 indicate high levels of structural conservation, and values greater than 1 imply compensatory substitutions in the base pairs. The alignment of an RNA gene can be evaluated with various measures for the evolutionary conservation of RNA structures, and although SCI is one of the simplest methods, it is also one of the most accurate (Gruber *et al.*, 2008).

The pipeline I have developed retrieves EPO and MultiZ genomic alignments, and provides the option to realign with MAFFT/X-INS-i and PicXAA-R, for comparison with the genomic alignments. A range of alignment statistics, including SCI, are calculated, and the presence of protein-coding genes and other RNA genes in the flanking regions is determined. These statistics can be used to either remove problematic sequences from an alignment, or to discard low quality alignments. I apply the pipeline to retrieve human RNA sequences from Rfam, retrieve EPO-12, EPO-35 and MultiZ genomic alignments, and filter them to produce high quality datasets of mammalian RNA gene alignments in an annotated genomic context.

2.1.3 The MARMOSET pipeline

There are several advantages to developing a software pipeline to generate alignments of RNA genes. Such alignments have potential applications beyond the scope of this thesis, and a pipeline enables others to generate new datasets, for different species or when the underlying data is updated. Moreover, the data processing is complex and

it is natural to split the work into discrete modules; this also makes it easy to apply different filters to the same source data, and to omit certain stages. The pipeline that produces RNA gene alignments is named **MARMOSET: Multiple Alignments of RNA for Models Of Structure and Evolutionary Theory**. The different stages of the MARMOSET pipeline often require similar code for tasks such as sequence manipulation and tree parsing, so I wrote a software library named **MonkeyShines** to provide this functionality.

2.2 Materials and methods

2.2.1 MonkeyShines software library

The MonkeyShines software library consists of Perl modules for general programming and bioinformatics tasks, and a module specific to the MARMOSET pipeline. (It also contains a module relating to the TARSIER pipeline, for evaluating RNA gene prediction, but description of that is deferred to a later chapter.) In some cases there is existing code for the functions in these modules, but if so it tends to be straightforward to write: for example, calculating the mean of an array of numbers or parsing a FASTA-format file. There are many freely available Perl libraries, including BioPerl for bioinformatics work, but these often provide too much functionality (making the code unnecessarily complicated), or not quite the desired functionality (meaning that edits or re-writes of existing code are required). The MonkeyShines library is available under an open access (GPL3) licence, at http://bitbucket.org/james_monkeyshines/monkeyshines/overview, and the functions are documented, so I will not describe the code in exhaustive detail. Rather, to give an idea of what the code is doing I will describe a representative function from each module, with the exception of the ‘Baboon’ module, which performs a novel alignment visualisation and so warrants a full explanation.

2.2.1.1 MonkeyShines Perl modules and example functions

`MonkeyShines::Sequence` The `excise_columns` function takes a sequence alignment and removes columns which match a criterion specified by a regular expression; the default is to remove columns that consist entirely of gap characters. Columns of gaps can occur, for example, in the MARMOSET pipeline when sequences that fail a quality check are removed.

`MonkeyShines::Tree` The `unroot` function takes a rooted tree and creates a trifurcation at the root, effectively unrooting it, in the absence of an explicit outgroup. If the two children of the root are internal nodes, then the choice of which one to collapse, to create the trifurcation, is arbitrary; this function does it with the first child it processes, which is the one that is leftmost in the original tree. The MARMOSSET pipeline uses this function to create acceptable trees for other software programs, such as PhyML and PHASE.

`MonkeyShines::Utils` The `coord_overlap` function takes two start and stop coordinates and determines whether they overlap, and if so, the nature of the overlap. The MARMOSSET pipeline uses this function to detect whether a protein-coding gene overlaps an RNA gene, and if so, whether the RNA gene is contained within an intron or whether it overlaps a boundary of a UTR or CDS region.

`MonkeyShines::Marmoset` The `crop_alignment` function takes an alignment of an RNA gene plus flanking sequence, and trims the flanking sequences to a fixed number of bases. This is necessary in the MARMOSSET pipeline because the EPO alignment is retrieved based on the genomic location of a reference RNA sequence and a given amount of flanking sequence. Gaps are invariably introduced when this sequence is aligned, and thus the lengths of the flanking regions will vary between alignments. This variation complicates later analyses, so the alignment is trimmed to restore a fixed flanking length on either side, including gaps in the reference sequence.

2.2.1.2 Visualising alignments with BABOON

The `MonkeyShines::Baboon` module is used for visualising alignments, with the name representing a rather tortuous backronym: **B**lurred **A**lignment **B**arcodes **O**f amino acids **O**r **N**ucleotides. During the course of developing the MARMOSSET pipeline I often wanted a quick visual overview of the genomic alignments that I was retrieving, so as to assess their quality at a glance. This is no substitute for a more thorough quantitative assessment, but it highlights the sorts of alignment properties that will be important to quantify. There are many programs for alignment visualisation, none of which produced the images I wanted in a scriptable manner. Jalview (Waterhouse *et al.*, 2009) is widely used, and the Mesquite package (Maddison and Maddison, 2011) is highly customisable, but both are standalone GUI programs that require several steps to generate an image from an alignment file. TEXshade (Beitz, 2000) is a powerful package

for generating high-quality alignment images, but is complex and has large memory requirements for long alignments. The key novelty of BABOON is the blurring that, perhaps counterintuitively, makes the alignments more comprehensible. Taylor (1997) described a systematic method for colouring amino acids based on their properties, such that blurring would produce sensible results. For example, hydrophobic amino acids are shades of green, so if a column consisting of alanine and leucine was blurred, the result would be a green column, emphasising that a hydrophobic residue is generally seen at that position while reducing visual complexity. If there were also hydrophilic amino acids at that position, the column would be more brown, as other colours would pollute the greenness. Although many existing visualisation tools implement the Taylor colour scheme, no software implements the blurring technique.

Alignments of nucleotides, rather than amino acids, are the focus of my work, but the same methods can be applied to generate digestible images. With nucleotide alignments the key properties are whether a column is GC-rich or AT-rich, or whether a column contains purines (R) or pyrimidines (Y). With these in mind, a systematic colouring similar to the Taylor scheme is possible, as shown in Table 2.1. Gaps are represented as white, and unknown characters are black, so a surfeit of these produce lighter and darker regions, respectively. It is easy to update the BABOON software to use any other colour scheme, e.g. if the distinction between *Keto* and *aMino* nucleotides is important.

BABOON permits blurring of rows instead of, or as well as, columns. In the amino acid case this would, for example, usefully highlight the hydrophobic and hydrophilic regions of a transmembrane protein. In the nucleotide case, GC-content is more apparent with horizontal blurring, and in both cases misalignments can become more obvious. To blur vertically, each position in a column is replaced by the colour derived from the mean of the R, G and B values at the sites in that column. To blur horizontally, the colour at a given position is replaced by the average of the R, G and B values at that site and those of neighbouring sites; the number of neighbours on each side is a user-defined parameter. At the extreme ends of the sequences there are fewer neighbouring sites, so the software uses as many neighbours as are available, but always equal numbers on either side. Blurring in both directions is achieved by blurring horizontally, and then vertically.

The BABOON software produces images in PNG format, and there are various display options, such as adding a legend for the colour scheme or marking a region of interest (e.g. an RNA gene).

Table 2.1: Nucleotide colour scheme.













Nucleotide	Colour	RGB Value	Swatch
A	Dark blue	(0, 0, 255)	
C	Yellow	(255, 255, 0)	
G	Red	(255, 0, 0)	
T (or U)	Light blue	(0, 255, 255)	
R (A or G)	Purple	(127, 0, 127)	
Y (C or T)	Green	(127, 255, 127)	
S (C or G)	Orange	(255, 127, 0)	
W (A or T)	Blue	(0, 127, 255)	
K (G or T)	Grey	(127, 127, 127)	
M (A or C)	Grey	(127, 127, 127)	
N	Black	(0, 0, 0)	
Gap (- or .)	White	(255, 255, 255)	

Figure 2.1 shows an amino acid alignment of a transmembrane protein (Or9a) from *Drosophila melanogaster* and three closely related fly species (Chang *et al.*, 2012). The transmembrane helices that Chang *et al.* annotate (which they say are consistent with other studies) clearly coincide with the colouring in Figure 2.1. Figure 2.2 shows an alignment of a snoRNA (RF01291) with 400 bases of flanking sequence on either side, for ten species from the EPO mammalian dataset. Note that both figures display the sequences on a single line, rather than the several wrapped lines that would be necessary if the sequence characters were displayed, enabling easier interpretation of the different regions of the alignments. It is also worth noting that the blurred and barcoded images are aesthetically pleasing, which may be of some consolation if one has to study a large number of them.

2.2.2 MARMOSSET pipeline structure

Figure 2.3 shows the structure of the MARMOSSET pipeline. The stages of the pipeline correspond to Perl scripts that are stored in the *marmoset* subdirectory of the MonkeyShines software package, which also contains instructions on how to execute the scripts and recommended values for filtering. In this section I describe each stage of the

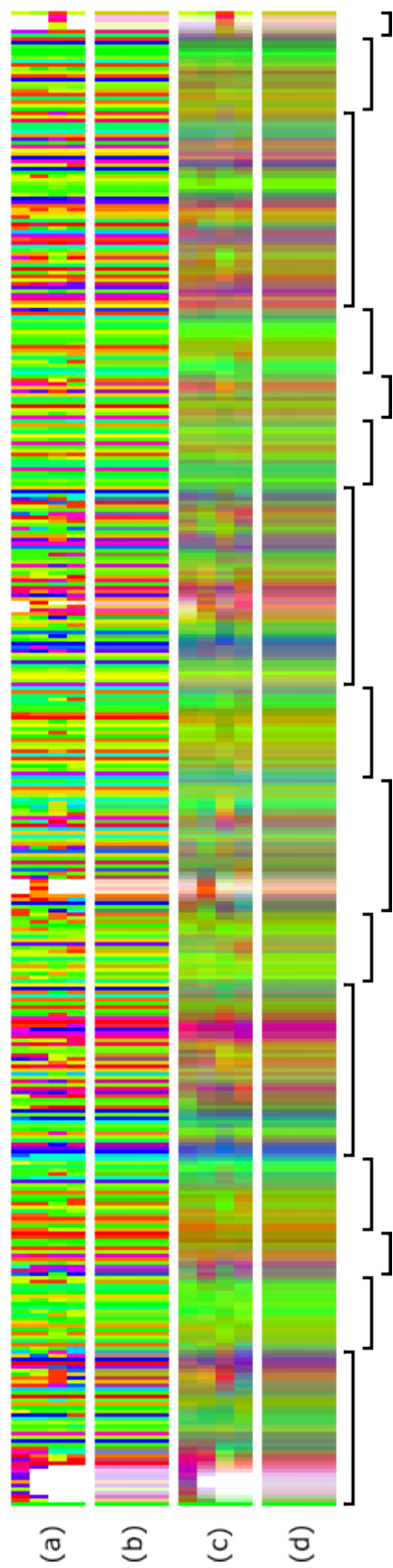


Figure 2.1: A four-species alignment of a transmembrane protein. Amino acids are represented as 2x10 pixel blocks of colour, according to the Taylor scheme, in which hydrophobic amino acids are green, hydrophilic are purple, and neutral are red. The black bars at the bottom of the figure indicate interior loops, helices, and exterior loops, in the top, middle, and bottom rows, respectively. Four degrees of blurring are shown: **(a)** an unblurred image: gaps and conserved columns are clear; **(b)** a 'barcode' image, blurred vertically: more vibrant colours indicate conservation, without the overwhelming detail of the unblurred image (the effect is more marked with more species); **(c)** a horizontally blurred image, averaged across 5 amino acids: larger scale patterns are more apparent, and differences between the sequences are highlighted; **(d)** a blurred barcode, in which the image in (c) is blurred vertically: patches of colour coincide with the structural annotation. Apart from the lettering of the images, the figure was wholly generated by the BABOON software.

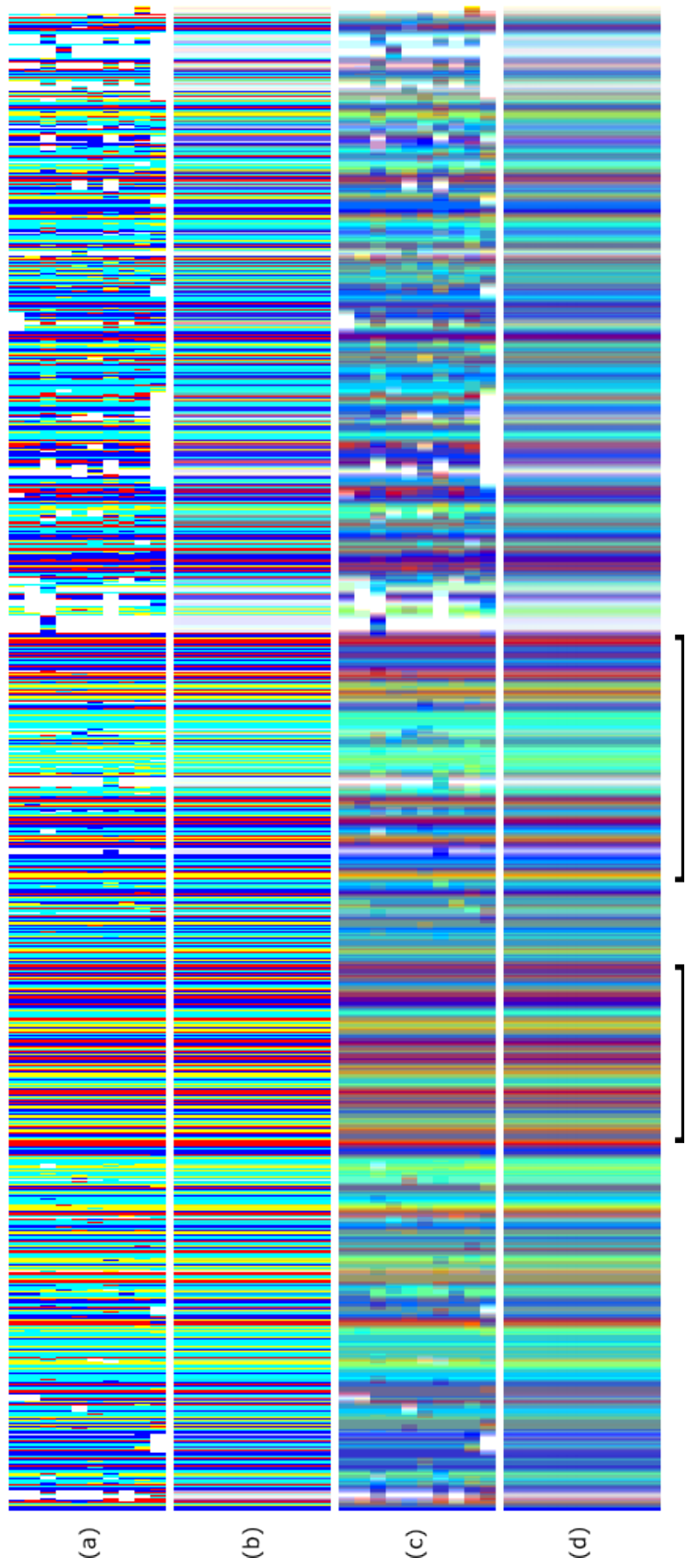


Figure 2.2: A ten-species alignment of an intronic snoRNA and flanking sequence. Nucleotides are coloured according to the scheme in Table 2.1. The black bars at the bottom of the figure indicate, from left to right, an exon (of the gene EIF4G2) and the snoRNA (RF01291). Four degrees of blurring are shown: **(a)** an unblurred image: gaps and conserved columns are clear; **(b)** a 'barcode' image, blurred vertically: patterns of AT-rich (blue) and GC-rich (red/orange/yellow) regions are obvious; **(c)** a horizontally blurred image, averaged across 5 bases: differences in the level of conservation between the exon and the RNA gene are clear; **(d)** a blurred barcode, in which the image in (c) is blurred vertically. Apart from the lettering of the images, the figure was wholly generated by the BABOON software.

pipeline in general terms, and the subsequent Results section describes an application of the pipeline to generate the datasets that are used in later chapters.

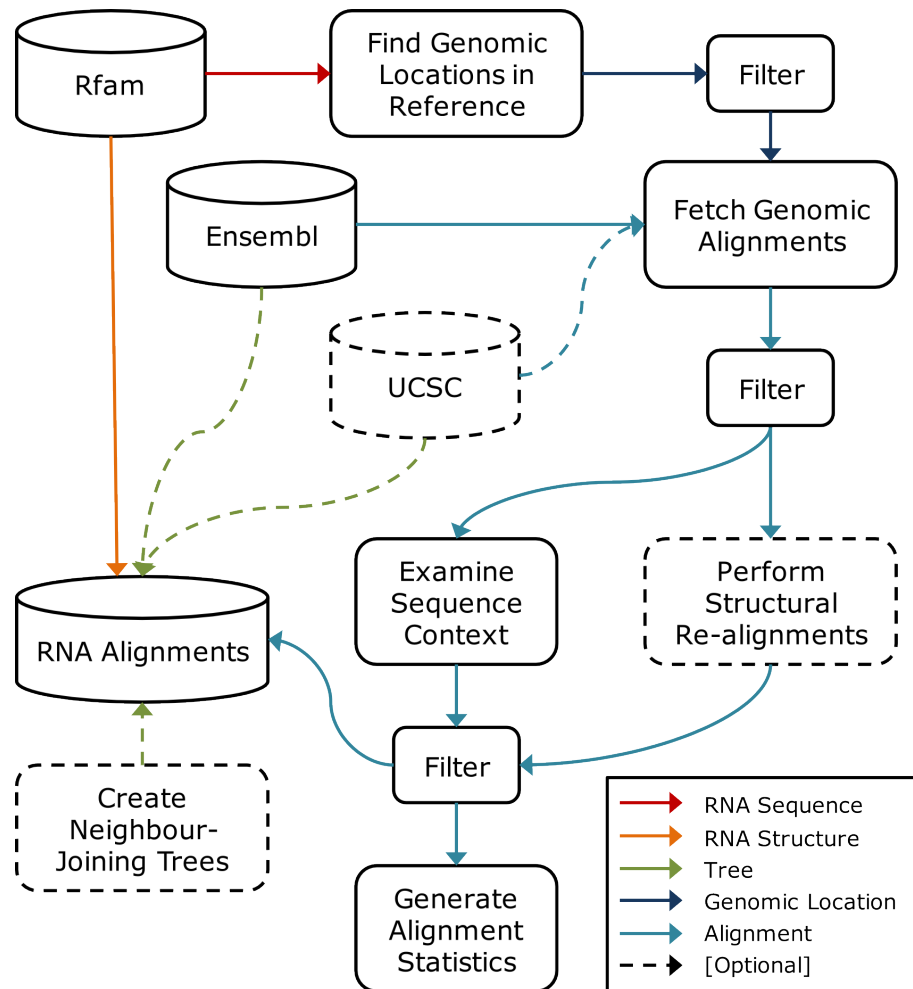


Figure 2.3: MARMOSSET pipeline structure. Cylinders represent external data sources, boxes are stages of the pipeline, and arrows represent different types of data, defined in the key at the bottom-right of the figure. Dashed lines indicate sources, stages, or data that are optional. The ‘Filter’ stages are deliberately ambiguous, as the filtering criteria will change depending on the scientific questions that one hopes to address with the final set of genomic alignments.

2.2.2.1 Rfam data

RNA alignments and structures are downloaded in flat files from Rfam (from the ‘seed’ dataset), and unnecessary data is removed to reduce them to a manageable size. The RNA sequences for a given reference species are extracted from the Rfam alignments, gaps are removed, and redundant sequences are flagged so that they can be filtered

out. The consensus structure for each alignment is also extracted, and mapped from the gapped reference sequence in the Rfam structural alignment to the ungapped RNA sequence. Some structures have paired bases at gap positions (because they have been annotated with respect to species other than the reference), and these are not used to avoid complications from having to edit the structures.

2.2.2.2 Find genomic locations in reference

The genomic location of an RNA sequence is determined with a BLAT search (Kent, 2002) against the genome of the reference species, and the properties of the best-scoring hit(s) are summarised to enable the removal of low-quality or ambiguous results. The BLAT score indicates how much of the query sequence was matched, and if the query and result sequences are identical this is noted as a ‘perfect’ match. A hit that maps to discontinuous genome sequence may cause problems for downstream analysis, so this is flagged, as are multiple non-overlapping hits with the same score, since duplications may also be problematic. It is also possible for one RNA query sequence to be a partial match to another query sequence, such that two distinct sequences map to the same genomic location, so this information is also included in the summary of the BLAT results.

2.2.2.3 Fetch genomic alignments

The Ensembl Perl API (Flicek *et al.*, 2012) makes it far easier to programmatically retrieve an EPO alignment for a particular genomic location than an equivalent MultiZ alignment, for which no such access is available (there is an API associated with the Galaxy project (Blankenberg *et al.*, 2010), but it does not seem to be actively developed and has no documentation). As mentioned in the introduction, EPO alignments also contain sequences from more recent assemblies than MultiZ alignments, so these are considered to be the main source for genomic alignments in the MARMOSSET pipeline. However, as MultiZ alignments have been widely used, it is useful to be able to retrieve these too, and this forms an optional stage in the pipeline.

Given a set of genomic locations of RNA genes, the EPO alignments are retrieved with the Ensembl Perl API, with a fixed amount of flanking sequence on either side (400 bases by default). Any EPO alignments that span multiple genomic blocks are discarded, to avoid the complication of joining blocks together. Occasionally the requested amount of flanking sequence is unavailable, and these alignments are also rejected. The gap percentage of each sequence is calculated, and if it is above a certain

threshold (25% of the gapped reference sequence length by default) the sequence is removed from the alignment. Long insertions that occur in a single species suggest a poor quality alignment and can cause problems for later analyses, so sequences with such inserts longer than a threshold (10% of the ungapped reference sequence length by default) are deleted. Ambiguous nucleotides ('N's) may indicate a poor quality region of the underlying genome assembly, and also cause a sequence to be removed from the alignment. If any of these criteria result in the removal of the reference species, then the whole alignment is discarded. This stage of the pipeline also calculates if the sequences of the RNA region are identical between species, and whether the RNA region is entirely absent in any species, as this information will be pertinent for downstream analyses.

Since there is no programmatic way to access MultiZ alignments, the MAF-format files are downloaded from the UCSC Genome Browser and processed locally. This processing requires the installation of the `bx-python` (http://bitbucket.org/james_taylor/bx-python) and `Galaxy` (<http://wiki.galaxyproject.org/Admin/GetGalaxy>: Blankenberg *et al.*, 2010) libraries, which provide functionality for indexing and manipulating the MAF files. As with the EPO alignments, flanking is added to the genomic locations of RNA genes, and the MultiZ alignments are extracted, either for all of the species in the dataset, or a subset of species.

It will almost always be necessary to join multiple MAF blocks, but the `Galaxy` scripts to join blocks take a very conservative approach and remove any species that are not common to all blocks, and this frequently leads to alignments of just 2 or 3 species. Instead, the `MARMOSET` pipeline joins blocks that are contiguous in the human reference sequence, only removing species if blocks contain sequences from different top-level scaffolds. This is rather unsophisticated, but subsequent filtering steps are effective at removing misaligned sequences and identifying low quality alignments. Apart from retaining alignments that span multiple blocks, the same processing steps that are applied to EPO alignments are applied to the joined MultiZ alignments.

2.2.2.4 Examine sequence context

Non-coding RNA molecules often regulate protein-coding genes, so it is interesting to know whether the RNA genes in the alignments are within the bounds of such a gene. If so, the RNA genes tend to be wholly within introns or UTRs, but some overlap CDS regions as a consequence of misannotation or due to some complicated biology.

The location of protein-coding genes is determined by the annotation of the human sequence in Ensembl (release 67). The ‘canonical’ transcript, which is usually the one with the longest protein product, defines a set of exons that are split into UTR and CDS regions if necessary. The location of these structures is compared to the bounds of the RNA gene, which is classified as either within or overlapping a CDS, an intron or a UTR. If none of these occur, the protein-coding gene is classified as either upstream or downstream, relative to the RNA gene.

In addition to protein-coding genes, there may be RNA genes in flanking regions, not only by chance but because some classes of RNA gene are often clustered on the genome. The MARMOSET pipeline detects RNA genes in flanking regions by submitting the human sequences for those regions to a homology search with Rfam, which runs a BLAST search as filtering step and then scans against any matching covariance models.

2.2.2.5 Perform structural re-alignments

The MARMOSET pipeline can re-align genomic alignments with two programs specifically designed for RNA alignment MAFFT/X-INS-I and PicXAA-R.

2.2.2.6 Generate alignment statistics

A range of common statistics are calculated for each alignment, including gap percentage, GC content, and mean pairwise identity (MPI). The MARMOSET pipeline also uses the structure associated with each alignment to calculate RNA-specific properties, such as SCI and the number and type (Watson-Crick, wobble, or mismatch) of paired bases in the alignment. Alignment visualisations are generated with the BABOON software described above, by default using 2 neighbours for horizontal blurring.

2.2.2.7 Create neighbour-joining trees

Some analyses in subsequent chapters require a fixed tree, and my original intention was to use the species tree provided by Ensembl that accompanies the EPO alignments. (The MultiZ phylogeny has an identical topology for the species that EPO and MultiZ have in common, and very similar branch lengths.) However, after performing Bayesian MCMC tree search on a small sample dataset, some consensus trees were significantly different from the species tree (this result is described and discussed in a later chapter). As an alternative to the species tree, the MARMOSET pipeline optionally

calculates a neighbour-joining (BIONJ: Gascuel, 1997) tree for each alignment, using PhyML 3.0 (Guindon *et al.*, 2010).

2.2.3 Source code

The MonkeyShines software library, incorporating the Marmoset pipeline, is freely available under an open-access licence at http://bitbucket.org/james_monkeyshines/monkeyshines/overview.

2.3 Results and discussion

2.3.1 MARMOSET pipeline execution

2.3.1.1 Rfam data

Human is used as a reference species, and I extract all human RNA sequences from the alignments in version 10.1 of the Rfam ‘seed’ dataset, a total of 1,403 sequences that are associated with 550 Rfam families (Gardner *et al.*, 2011). There are 148 duplicate sequences and 194 sequences in which the structure cannot be mapped, and these are removed to leave 1,061 RNA sequence spanning 524 families.

2.3.1.2 Find genomic locations in reference

The 1,061 RNA sequences from the previous step are submitted to a BLAT search against version GRCh37/hg19 of the human genome, and taking the best-scoring hit(s) gives a list of 1,210 results. I discard those that return hits that map to discontinuous genome sequence. Non-perfect matches and hits on the mitochondrial chromosome are filtered out, and if two different RNA sequences have overlapping genomic locations only the longest is retained. If there are two or more non-overlapping results with the same score, these are all discarded, since they may represent paralogous RNA genes, which make it harder to interpret the results in the later chapter on model selection. Table 2.2 shows the effect of these filters on the size of the dataset, which leave a set of 858 non-overlapping sections of the human genome, corresponding to 480 Rfam families.

Table 2.2: Results of filtering human genomic locations.

	Rfam Families		Genomic Locations	
	Adjustment	Total	Adjustment	Total
<i>Pre-Filtering Total</i>		524		1210
Zero BLAT Hits	-9	515	-22	1188
Discontiguous Genome Sequence	-11	504	-50	1138
Non-Perfect Hits	-16	488	-81	1057
Overlapping Genomic Locations	-1	487	-21	1036
Mitochondrial Chromosome	0	487	-12	1024
Multiple Equally-Good BLAT Hits	-7	480	-166	858
<i>Post-Filtering Total</i>		480		858

2.3.1.3 Fetch genomic alignments

I retrieve EPO-12 and EPO-35 mammalian alignments from Ensembl, which cover the species and assemblies shown in Table 2.3. I also retrieve MultiZ alignments for the 11 species that are shared between the EPO-12 and MultiZ alignments (pig is not included in the MultiZ dataset). The results of filtering on sequence and alignment quality, using the default settings described in the Materials and Methods section, are shown in the top section of Table 2.4. In order to provide subsequent analyses with adequate data, alignments with fewer than 5 distinct sequences in the RNA region are removed from the datasets.

Some examples of the consequences of filtering are given in Figure 2.4. Figure 2.4a shows an EPO-12 alignment before filtering; the sequence with ambiguous bases (shown in black) is removed, followed by the species with the long inserts that are not shared by the remaining species, resulting in the alignment shown in Figure 2.4b. It is uncommon to encounter alignments in which the RNA region is absent (Figure 2.4c), which suggest that the RNA gene has been gained or lost in some species; these alignments are removed because of their uncertain effect on downstream analyses. Some RNA genes are very well conserved at the sequence level, and Figure 2.4d shows an extreme example of a miRNA that is perfectly conserved in primates; this effectively only represents one sequence, and is deleted from the dataset.

Table 2.3: Species in EPO alignments (Ensembl release 67).

Common Name	Species	Assembly	Coverage	EPO-12
Human	<i>Homo sapiens</i>	GRCh37	High	✓
Chimpanzee	<i>Pan troglodytes</i>	CHIMP2.1.4	High	✓
Gorilla	<i>Gorilla gorilla</i>	gorGor3.1	Low	✓
Orangutan	<i>Pongo abelii</i>	PPYG2	High	✓
Macaque	<i>Macaca mulatta</i>	MMUL 1.0	High	✓
Marmoset	<i>Callithrix jacchus</i>	C_jacchus3.2.1	High	✓
Mouse	<i>Mus musculus</i>	NCBIM37	High	✓
Rat	<i>Rattus norvegicus</i>	RGSC 3.4	High	✓
Cow	<i>Bos taurus</i>	UMD3.1	High	✓
Pig	<i>Sus scrofa</i>	Sscrofa10.2	High	✓
Horse	<i>Equus caballus</i>	EquCab2	High	✓
Dog	<i>Canis familiaris</i>	CanFam 2.0	High	✓
Gibbon	<i>Nomascus leucogenys</i>	Nleu1.0	High	
Tarsier	<i>Tarsius syrichta</i>	tarSyr1	Low	
Mouse Lemur	<i>Microcebus murinus</i>	micMur1	Low	
Bushbaby	<i>Otolemur garnettii</i>	OtoGar3	High	
Tree Shrew	<i>Tupaia belangeri</i>	tupBel1	Low	
Kangaroo Rat	<i>Dipodomys ordii</i>	dipOrd1	Low	
Guinea Pig	<i>Cavia porcellus</i>	cavPor3	High	
Squirrel	<i>Ictidomys tridecemlineatus</i>	spetri2	High	
Rabbit	<i>Oryctolagus cuniculus</i>	oryCun2	High	
Pika	<i>Ochotona princeps</i>	OchPri2.0	Low	
Alpaca	<i>Vicugna pacos</i>	vicPac1	Low	
Dolphin	<i>Tursiops truncatus</i>	turTru1	Low	
Cat	<i>Felis catus</i>	CAT	Low	
Panda	<i>Ailuropoda melanoleuca</i>	ailMel1	High	
Microbat	<i>Myotis lucifugus</i>	Myoluc2.0	High	
Megabat	<i>Pteropus vampyrus</i>	pteVam1	Low	
Hedgehog	<i>Erinaceus europaeus</i>	eriEur1	Low	
Shrew	<i>Sorex araneus</i>	sorAra1	Low	
Elephant	<i>Loxodonta africana</i>	Loxafr3.0	High	
Hyrax	<i>Procavia capensis</i>	proCap1	Low	
Hedgehog Tenrec	<i>Echinops telfairi</i>	TENREC	Low	
Armadillo	<i>Dasypus novemcinctus</i>	dasNov2	Low	
Sloth	<i>Choloepus hoffmanni</i>	choHof1	Low	

Table 2.4: Results of filtering genomic alignments.

	EPO-12			EPO-35			Multiz		
	Rfam Families	Genomic Locations		Rfam Families	Genomic Locations		Rfam Families	Genomic Locations	
<i>Pre-Filtering Total</i>	480	858		480	858		480	858	
No Genomic Blocks	-5	475	-27	-5	475	-25	0	475	-1
Multiple Genomic Blocks	-29	446	-48	-29	446	-48	N/A	N/A	N/A
Ambiguous Bases (Human)	0	446	0	0	446	0	0	446	0
Long Inserts (Human)	-14	432	-32	-5	441	-17	-10	470	-21
More than 25% Gaps (Human)	0	432	0	0	441	0	0	441	0
Insufficient Flanking	0	432	-2	-1	440	-3	0	440	0
Gene Gain/Loss	-2	430	-3	-1	439	-3	-1	469	-6
Fewer than 5 Sequences	-277	153	-563	-90	349	-270	-234	235	-531
CDS Overlap	-11	142	-16	-20	329	-29	-18	217	-27
UTR/Intron Overlap	-3	139	-3	-6	323	-7	-5	212	-5
RNA in Flanking	-26	113	-27	-67	256	-86	-41	171	-47
SCI < 0.8	-6	107	-13	-53	203	-83	-22	149	-38
<i>Post-Filtering Total</i>	107	124	124	203	287	287	149	182	182

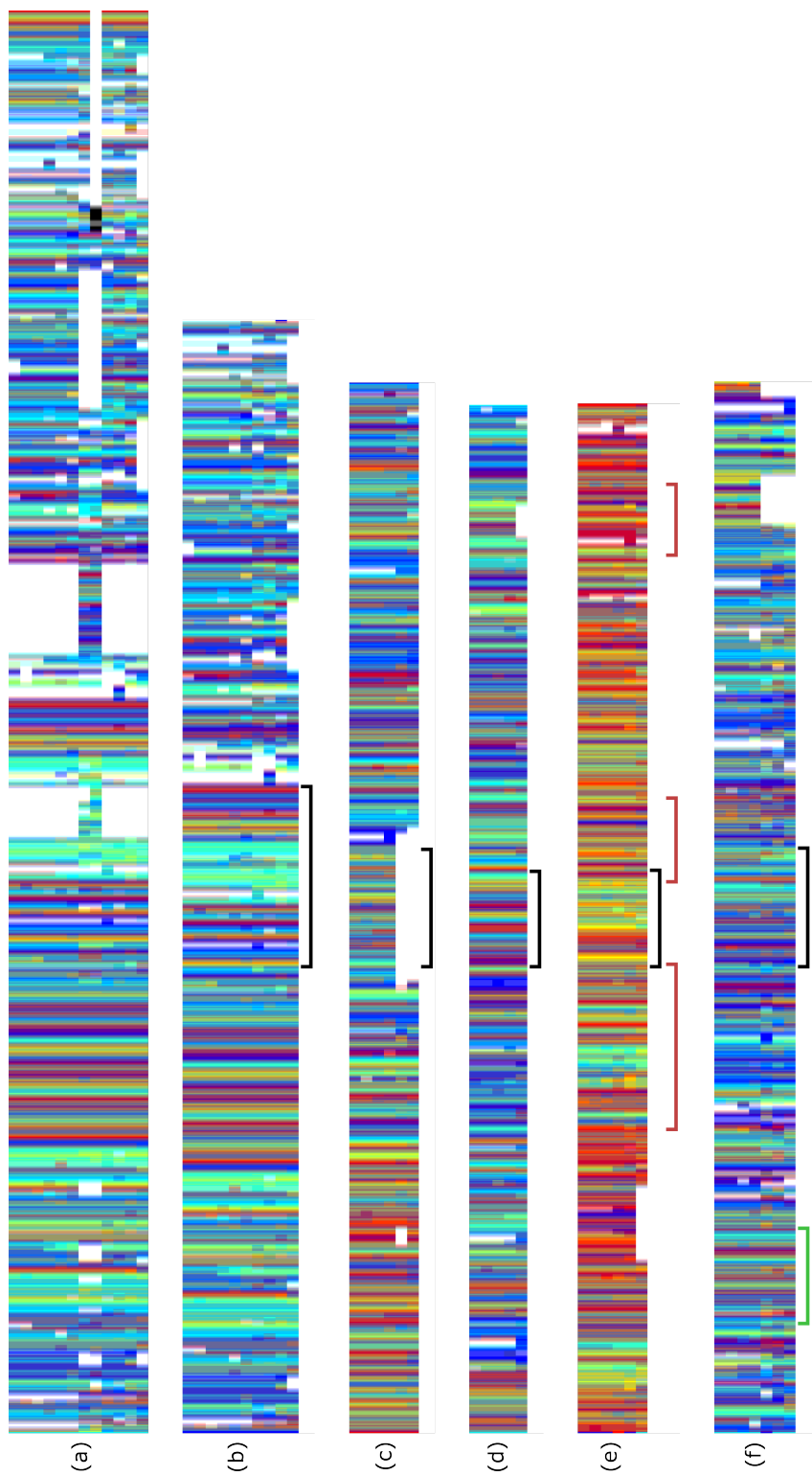


Figure 2.4: Examples of EPO-12 alignments of RNA genes with 400 flanking bases, coloured and blurred with the default BABOON settings described in section 2.2.1.2. **(a)** A snoRNA (RF01291) before filtering to remove sequences with ambiguous bases and long insert regions. **(b)** The snoRNA in (a) after filtering, with the RNA region marked by a black bracket. **(c)** A snoRNA (RF01210) that is only present in great apes. **(d)** A perfectly conserved miRNA precursor (RF00027). **(e)** A pre-miRNA (RF01039) that overlaps the CDS regions (red brackets) of the canonical transcript of the human gene SCRIB. **(f)** A pre-miRNA (RF00130) with another RNA gene (green bracket) in the flanking region.

2.3.1.4 Examine sequence context

It is useful to know whether an RNA gene lies within a protein-coding region when interpreting subsequent analyses, and overlapping annotation of CDS regions and RNA genes, as in Figure 2.4e, can be identified and removed, since they make reasoning about evolutionary events difficult. Alignments in which RNA is detected in the flanking region (e.g. Figure 2.4f) are filtered out, for two reasons. Firstly, these clusters of RNA genes are typically from the same family, and thus will have similar sequences, which could lead to the alignment of paralogs rather than orthologs. Secondly, it greatly simplifies the evaluation of RNA gene prediction software if it is known that there are no RNA signatures in the flanking sequence which could lead the software astray.

2.3.1.5 Generate alignment statistics

The SCI value is used to judge whether the RNA region of an alignment resembles a structurally conserved RNA gene, and a cut-off value of 0.8 is used. This value is somewhat arbitrary, but is effective at removing alignments that do not match the associated secondary structure, which may mislead subsequent inference.

2.3.2 RNA gene properties

The MARMOSSET pipeline tends to produce datasets with particular properties, due to biases in the Rfam source data and to the filtering steps of the pipeline. Table 2.5 shows the numbers of different RNA types in the original and filtered datasets. I use the Rfam ‘seed’ dataset, to obtain the most reliable set of sequences and structures, but this does mean that there are few genes in some notable categories, such as rRNA and tRNA. However, these RNA genes may be the exception rather than the rule with respect to structural and evolutionary patterns, due to their fundamental importance to all life and their subsequent conservation. In this thesis, I use the dataset to evaluate *de novo* gene prediction and to assess the fit of different models of evolution, and a lack of well known types of RNA genes is not necessarily problematic for these aims. There are already programs to predict well known types of RNA gene (e.g. RNAMMER Lagesen *et al.*, 2007, tRNAScan-SE Lowe and Eddy, 1997), so while gene prediction software should find such genes, they will generally not be used to do so. It has been shown several times that RNA models better describe evolution in rRNA genes than DNA models (refer to the introduction to Chapter 5 for further details), and my intention in this thesis is to examine whether this holds for other types of RNA gene. Using the Rfam ‘seed’ dataset gives a set of RNA genes that represent the current state of knowledge in this area, and although this is certainly biased, it is a good starting point; and by constructing a pipeline to gather the alignments I hope to have made it easy to perform similar analyses in the future, as the suite of RNA genes in Rfam expands.

Table 2.5: RNA types in the pre-filtered dataset, and in the filtered datasets for each set of genomic alignments. The ‘Other’ type is a heterogeneous mixture of molecules such as cis-regulatory elements and selenocysteine insertion sequences that do not naturally fit into other groups.

RNA Type	Pre-filtering	Post-filtering		
		EPO-12	EPO-35	MultiZ
Long ncRNA	111	19	24	26
microRNA	256	34	106	50
Ribosomal	1	0	1	0
RNase P	1	0	1	0
scaRNA	22	6	11	5
snoRNA	346	54	119	84
Spliceosomal	1	0	1	0
tRNA	3	0	1	1
Vault	2	0	1	0
Other	115	11	22	16

Many of the known vertebrate RNA genes are miRNA and snoRNA, which are relatively short; in the filtered EPO-35 dataset the minimum and maximum lengths of these two types are 61 and 420, respectively, with a median length of 96 bases. In Rfam, conserved regions of long ncRNA genes, rather than the entire RNA gene, are often stored as distinct entries, and the length of the ‘genes’ in this category is also relatively short, with a median value of 133 bases. Treating each of these regions as a gene is a simplifying assumption that should not affect the results too greatly, as each is defined as a domain with a secondary structure that should resemble a short RNA gene.

The proportion of paired bases in an alignment (calculated across all sequences, according to the consensus secondary structure) ranges from 8% to 95% in the EPO-35 dataset, and is dependent on the type of RNA gene. The miRNA genes have the highest proportion on average, with a mean value of 62%, and snoRNA and long ncRNA genes have means of 39% and 35%, respectively. Across all of the genes, most of the paired bases are Watson-Crick pairs (median 85%) or wobble pairs (median 12%), which is encouraging since it indicates that the secondary structures are plausible. The GC content of the RNA gene alignments is approximately normally distributed (Shapiro-Wilk test, $p=0.01$), with mean 0.48 and standard deviation 0.10.

The filtering steps in the MARMOSET pipeline remove different proportions of RNA genes from the original dataset depending on the alignments that are used. The filtered EPO-12 dataset retains 14% of the original RNA genes, compared to 33% for the EPO-35 dataset. The difference here is largely due to the constraint that alignments have at least 5 sequences; it is clearly going to help if you start with more species. The MultiZ dataset represents 21% of the original dataset, which is more than the EPO-12 dataset due to the necessarily absent filter on multiple alignment blocks.

Filtering is not generally biased towards RNA genes with a particular property. A disproportionate number of RNA genes in the ‘Other’ category (Table 2.5) are filtered, but since these are poorly defined in the first place this is neither too surprising nor too troubling. Since the proportion of paired bases is closely linked to the RNA type, the distribution of this property is similarly unaffected by filtering, and the GC content remains approximately normally distributed around the same mean value (0.48, SD 0.09).

2.3.3 Comparing alignments

2.3.3.1 EPO and MultiZ alignments

It is somewhat tricky to compare the genomic alignments produced by the EPO and MultiZ methods, as they cover different sets of species and different assemblies, but there are a couple of ways to compare them. The first is to generate summary statistics for each and compare the distribution of alignment properties, such as gap length or alignment length. The strict filtering of the original genomic alignments removes many of the large gaps, and the remaining gaps tend to be very short (Figure 2.5). The gap distribution differs between RNA gene regions and flanking regions, with the latter tending to be longer, as one would expect from less constrained regions. The gap distributions between different alignment sources are very similar, as are the distributions of alignment length (data not shown).

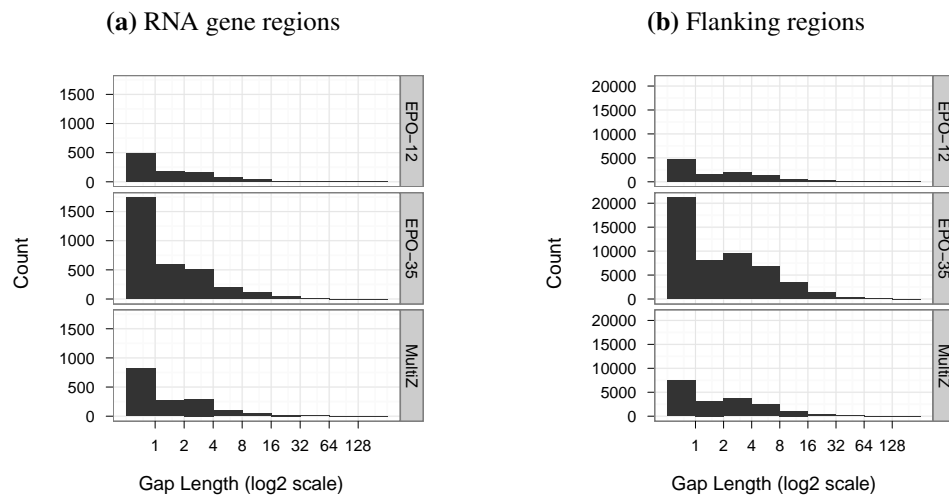


Figure 2.5: Gap length distributions for **(a)** RNA gene regions and **(b)** flanking regions. Most gaps are a single character in length, and there is a tendency for longer gaps in flanking regions.

Another means to compare the alignments is to compare the sequences themselves; the MultiZ alignments are a superset of the EPO alignments, because the filtering is more permissive for the MultiZ alignments in allowing multiple genomic blocks. For the alignments in both datasets, I took the sequences of the common species and calculated the Needleman-Wunsch distance to measure sequence similarity (Needleman and Wunsch, 1970), which is equivalent to the Levenshtein edit distance (Sellers, 1974). Calculations were done with ‘needle’ from the EMBOSS software package (Rice *et al.*, 2000). Figure 2.6 shows that, while there is no guarantee that the sequences represent the same genomic location, at the sequence level they are very similar, with only a handful of changes in the RNA region, and not many more when extending out to 400 bases flanking on either side. Given that the sequences are often from different assemblies, it is perhaps surprising that there are few large differences between the sequences from the different alignments. However, the RNA genes are often well-conserved and have been rigorously filtered, so the comparison between the EPO and MultiZ alignments is of well-aligned sequences, of relatively-easy-to-align regions. I would not necessarily expect the similarity to extend to regions of non-coding DNA in general.

2.3.3.2 Structural re-alignments

In contrast to the comparison of EPO and MultiZ alignments, it is possible to rigorously compare a genomic alignment and its structural realignment, as both contain exactly

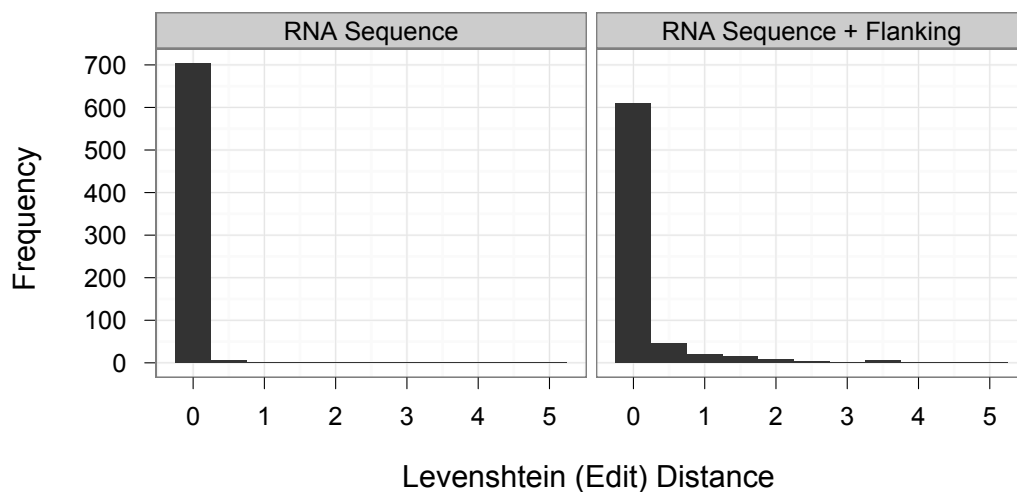


Figure 2.6: Pairwise difference between species common to EPO and MultiZ alignments of the same RNA sequence. The x-axis was truncated at 5. Four comparisons had values between 5 and 16, three of which were *Callithrix jacchus*, the remaining one being rat. *C. jacchus* has a new assembly in the EPO alignments (it is not assembled onto chromosomes in the MultiZ data), and this accounts for these differences. In the case of the rat comparison, both alignments use the same genomic location, but MultiZ aligns a small additional region at one end.

the same characters. The MetAl program (Blackburne and Whelan, 2012) calculates a distance metric between two alignments, and this is used to compare the two structural alignments (MAFFT and PicXAA-R), and EPO-12 and PicXAA-R. Both structural alignments produce very similar results, so EPO-12 and MAFFT are not compared. The default ‘d-pos’ metric of MetAl is used, which is equivalent to the probability that a random base would align to another location in a randomly selected sequence. Figure 2.7 demonstrates that the EPO genomic alignments are not very different from structural alignments, because the distance values are mostly close to zero. The structural alignments are not used in subsequent chapters, in order to simplify the interpretation of the results.

2.3.4 Conclusions

In this chapter I have outlined the need to generate a set of RNA gene alignments from genomic alignments, in order to study the properties of those genes from an evolutionary and phylogenetic perspective. I presented the MARMOSSET pipeline as a means to generate such alignments, and described a novel implementation of a colouring scheme

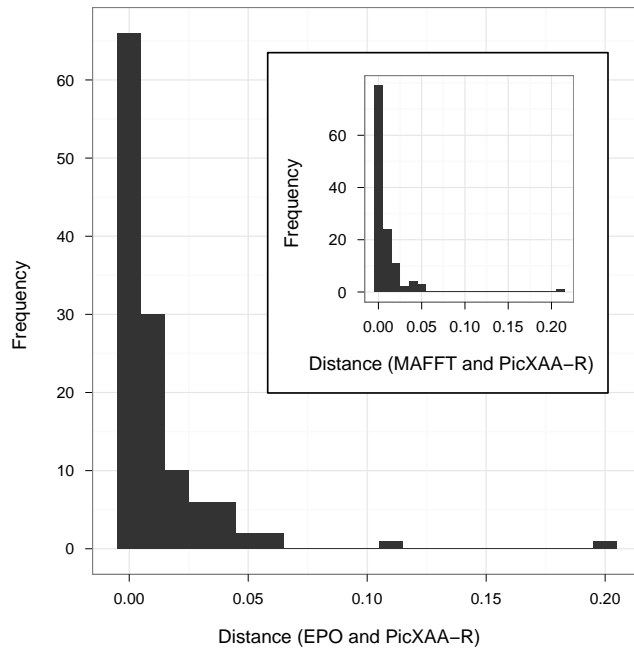


Figure 2.7: Distance between EPO genomic alignments and structural re-alignments, using the default ‘d-pos’ metric of MetAl (Blackburne and Whelan, 2012) which has a value between 0 and 1.

that creates useful visualisations of alignments. The EPO and MultiZ pipelines use different approaches to generate alignments, but the alignment properties are similar, and the sequences in common between the alignments are usually identical. Strict filtering can be attritional, causing a significant portion of the original alignments to be discarded, but the greater numbers of sequences in the EPO-35 alignments go some way to assuage this problem. Despite nominally including ‘low-coverage’ genomes, the EPO-35 alignments are of a similar quality to the EPO-12 alignments, and thus offer a viable alternative to the ‘high-coverage’ EPO-12 alignments. In the following chapters I use the filtered genomic alignments of RNA genes to evaluate *de novo* RNA gene prediction and to investigate the usefulness of RNA models of evolution.

Chapter 3

Evaluating *de novo* RNA gene prediction software

3.1 Introduction

De novo RNA gene prediction has the potential to uncover new families of ncRNA, in turn improving knowledge about the evolution of RNA genes, in turn hopefully leading to better methods of prediction. RNA gene prediction usually requires sequences from multiple species, because the thermodynamic stability of a single sequence is not generally distinguishable from that of suitably random sequences (Workman and Krogh, 1999; Rivas and Eddy, 2000); although this is not necessarily the case for certain types of small RNA, such as miRNA precursors (Bonnet *et al.*, 2004). Some programs for RNA gene prediction are limited to certain ncRNA families (reviewed in Machado-Lima *et al.*, 2008), and others are taxonomically restricted (e.g. Herbig and Nieselt, 2011), but I will concentrate on general methods that are applicable to large mammalian datasets. In the introductory chapter I briefly reviewed different methods of *de novo* RNA gene prediction, and in this introduction I provide more details about the programs, examples of their application, and evaluation of their accuracy.

3.1.1 *De novo* RNA gene prediction software

The DYNALIGN program, which simultaneously predicts structure and a pairwise alignment, has been adapted for RNA gene prediction (Uzilov *et al.*, 2006), as has the similar FOLDALIGN (Havgaard *et al.*, 2007). DYNALIGN and FOLDALIGN are computationally expensive, and are currently not practical for scanning genomes for RNA

genes; the use of pairwise rather than multiple alignments may also limit the performance of these methods, depending on the species being studied.

MSARI (Coventry *et al.*, 2004) uses thermodynamic information and a measure of covariation to predict RNA genes, and allows for a limited amount of misalignment, to reduce the dependence on the quality of the sequence alignment; it has now been superseded by more sophisticated software. RNAz is explicitly designed with the aim of fast execution, and uses multiple sequence alignments and a machine learning method that combines measures of evolutionary conservation and thermodynamic stability (Washietl *et al.*, 2005b; Gruber *et al.*, 2010).

QRNA (Rivas and Eddy, 2001) uses probabilistic models and has good accuracy, but its practicality is limited by slow execution and the restriction to pairwise alignments. In the EvoFold software, a phylo-grammar models both secondary structure and sequence evolution, and requires a tree with branch lengths to be provided (Pedersen *et al.*, 2006). Bradley *et al.* (2009b) develop and compare a range of phylo-grammar models, including EvoFold, but their sophisticated approach is correspondingly complex, and has not been widely applied. CMfinder (Yao *et al.*, 2006) predicts RNA motifs in unaligned homologous sequences with covariance models, and can be used for detection of RNA genes; heuristics that incorporate thermodynamic and comparative data are used to generate local alignments in which structural motifs are detected and then merged, to form putative RNA genes.

In this chapter I evaluate the accuracy of RNAz and EvoFold, so it is useful to describe their algorithms in further detail, in order to provide some context for the interpretation of my results.

3.1.1.1 RNAz

RNAz is based on a method of RNA secondary structure prediction that models the thermodynamics of RNA molecules. Stacking interactions between neighbouring base pairs in a secondary structure give helices stability, and can be used to calculate a good approximation of the free energy of the helix as a whole. Between or within helices, different types of single-stranded loop can occur, and these have less well-defined free energy calculations than helices, and are usually defined by loop length rather than base composition. Methods to predict the RNA structure based on minimum free energy (MFE) calculations have been widely used (e.g. Zuker and Stiegler, 1981; Mathews *et al.*, 2004; Markham and Zuker, 2008), although they can perform poorly for large molecules (Reeder *et al.*, 2006).

There are two versions of RNAz, both of which have the same methodology; the algorithm is more clearly explained in Gruber *et al.* (2010) than in Washietl *et al.* (2005b). RNAz measures two independent properties of an RNA structure, both of which are based on the thermodynamics of the molecule. The first of these is thermodynamic stability, which is judged by calculating z-scores to determine how much the MFE of a sequence (calculated with RNAfold (Hofacker *et al.*, 1994; Hofacker, 2003)) differs from randomised sequences with the same base composition and length. RNAz uses a machine learning technique (based on nucleotide composition) to estimate the required values for the randomisations, because exact calculations are computationally expensive for genome-wide analysis.

The second property that RNAz measures is structural conservation, which is evaluated using the Structural Conservation Index (SCI; described in section 2.1.2.3). The MFE of the whole alignment is calculated with RNAalifold (Hofacker *et al.*, 2002; Bernhart *et al.*, 2008), which incorporates information on covariation, and the ratio of this and the mean of the individual MFEs for each sequence gives the SCI. The z-scores and SCI values are used as input for another machine learning model, a support vector machine (SVM), which subsequently determines whether the input represents structural RNA. The SVM classifier also provides a “probability”, which is a number between 0 and 1 but is not a probability in the formal sense of the word.

In version 1.0 of RNAz the number of sequences and their mean pairwise identity (MPI) were also provided to the SVM classifier, which restricted the program to analysis of a maximum of 6 sequences. In version 2.0 these values are replaced by a normalised Shannon entropy value, which achieves much the same effect, and removes the limit on the number of sequences in the input alignments. Another improvement in version 2.0 is the use of a more sophisticated dinucleotide model for calculating z-scores, providing a better null model and thus better discrimination between RNA and other sequences.

Training alignments for the SVM classifier were generated in the same manner for both versions of RNAz (although the dataset was larger for version 2.0), following an approach that enabled the results to be compared with previous work (Washietl and Hofacker, 2004). Sequences from Rfam were clustered based on pairwise identity, then sets of sequences were randomly chosen from the clusters and aligned with ClustalW (Larkin *et al.*, 2007). In version 1.0 a randomised version of each alignment was created by a shuffling procedure that maintained some alignment properties, but not dinucleotide content. This was a weakness in the method (Babak *et al.*, 2007),

and in version 2.0 randomised alignments in the negative training set were created with either Multiperm (Anandam *et al.*, 2009) or SSSIz (Gesell and Washietl, 2008), depending on the information content of the alignment. Despite this improvement in the training for version 2.0, Gruber *et al.* (2010) highlight the problem of poor-quality alignments that do not account for RNA structure, and suggest that as more taxa are added, increasing evidence of conservation may be counteracted by alignment errors.

3.1.1.2 EvoFold

In contrast to the thermodynamic approach of RNAz, EvoFold uses phylo-grammars (see section 1.1.2) that explicitly model the evolution of sequence and structure with stochastic context-free grammars (SCFGs). EvoFold uses two phylo-grammars, one for regions with RNA genes, the other for “background” regions that lack RNA genes. The probabilities obtained when each is applied to an alignment are then used to determine whether it contains an RNA structure. The RNA phylo-grammar has a DNA model for loops and unpaired regions, and an RNA model for paired bases; the background phylo-grammar has the same DNA model (and no RNA model). The RNA model has 16 states, describing the changes between all dinucleotides, each of which has an associated equilibrium frequency (15 free parameters); there are 15 parameters to describe the rate of change between each combination of canonical base pairs, plus two further parameter for changes within mismatch pairs, and between mismatch and canonical pairs. The DNA model is derived from the RNA model using a marginalisation strategy intended to minimise unimportant differences between the models (supplementary information in Pedersen *et al.*, 2006).

EvoFold predicts RNA genes by means of the folding potential score (fps), defined as the log odds ratio of the likelihoods of observing an alignment, a , with the RNA (φ_{RNA}) and background ($\varphi_{background}$) phylo-grammars:

$fps = \log(P(a|\varphi_{RNA})/P(a|\varphi_{background}))$. The value is length-dependent, so a normalised version of the score is used in the paper, although the method of normalisation is not explicitly stated. The original version of the EvoFold software only calculates the probabilities of each phylo-grammar, requiring the user is to perform the log odds calculation and a sensible length-normalisation step. A later version of EvoFold, unpublished but available on request from the lead author, is more user-friendly, and outputs a comprehensive set of results including the fps and the regions of the alignment that are predicted to contain RNA structure.

The RNA phylo-grammar was trained by mapping all human Rfam sequences to

the human genome, retrieving corresponding sections of the 8-speciesMultiZ genomic alignment, and annotating each with the consensus structure from Rfam. These alignments were then filtered to remove nonconserved regions (according to the phastCons method, Siepel *et al.*, 2005), repeats, and regions that did not preserve synteny between human and mouse. The same phylogenetic tree was used for all training runs, calculated from the MultiZ genomic alignment by using phastCons.

3.1.2 Genome scans

The description of the RNAz software (Washietl *et al.*, 2005b) was followed by an application to a filtered subset of the human genome (Washietl *et al.*, 2005a), and a similar analysis was done with EvoFold (Pedersen *et al.*, 2006). Subsequently, the regions in the ENCODE pilot project (ENCODE Project Consortium, 2007) were used to compare the performance of RNAz, EvoFold and CMfinder (Washietl *et al.*, 2007; Torarinsson *et al.*, 2008), with the interesting result that the three programs detect minimally overlapping sets of RNA genes. These genome scans were conducted by the authors of the software in the analyses, and aimed to provide a measure of the accuracy of the programs, to justify their use by other researchers.

CMfinder is computationally demanding, and has generally only been used by the authors of the software, to analyse prokaryote genomes (Weinberg *et al.*, 2007; Yao *et al.*, 2007; Weinberg *et al.*, 2010). EvoFold and RNAz have been used to predict RNA genes in a wide range of organisms, including yeast (Steigele *et al.*, 2007), fruit flies (Rose *et al.*, 2007; Stark *et al.*, 2007; Zhong *et al.*, 2012), plants (Song *et al.*, 2009), and vertebrates (Parker *et al.*, 2011).

3.1.3 Evaluating RNA gene prediction

The evaluation of software for RNA gene prediction has, perhaps surprisingly given the popularity of RNAz and EvoFold, rarely been done in an independent and systematic manner. To calculate the accuracy of RNA gene prediction software it is necessary to provide sets of alignments both with and without RNA genes. As was demonstrated in Chapter 2, generating a positive dataset of RNA gene alignments is not trivial, and alignment quality can have a significant impact on the results (Saito *et al.*, 2010). A negative dataset is used to calculate the false positive rate, which is important because while it is relatively easy to experimentally verify the very best candidate RNA genes, these are only a small proportion of the total, and a low false positive rate is required

for confidence in the predictions.

In the evaluation of EvoFold and RNAz on the ENCODE regions (Washietl *et al.*, 2007), the estimates of the false positive rate are high, ranging from 50% to 70%, and these are likely to be underestimates, because appropriate negative datasets that preserve dinucleotide content were not used (Babak *et al.*, 2007). As Washietl *et al.* note, their study also suffers from a lack of known RNA genes in the ENCODE regions, providing a relatively sparse positive dataset that cannot provide a reliable true positive rate. Babak *et al.* (2007) tested a range of software, including RNAz and EvoFold, and found that results for most programs and RNA types correlated with sequence conservation, and that different types of RNA were differentially detected. The positive dataset in this case was gathered in a similar manner to that described in Chapter 2, albeit without flanking sequences. The negative dataset was carefully generated to preserve dinucleotide content, but unfortunately the exact shuffling algorithm limits the evaluation to pairwise alignments. This limitation inspired the creation of two programs that approximately preserve dinucleotide content (and other properties, such as gap percentage and local conservation patterns): SISSIz (Gesell and Washietl, 2008) uses a phylogenetic approach, and its methodology has been used to improve the accuracy of the latest version of RNAz (Gruber *et al.*, 2010); Multiperm (Anandam *et al.*, 2009) uses a graph theoretical method. Alternatively, sophisticated simulations of genomic background data are also able to effectively estimate false positive rates (Varadarajan *et al.*, 2008; Bradley *et al.*, 2009b).

The high rate of false positives seen in all genome scans means that candidate RNA genes are often processed to get a higher confidence dataset. Some genome scans do this in an *ad hoc* manner, according to the particular species under study (e.g. Stark *et al.*, 2007; Weinberg *et al.*, 2007), but LocARNA-P (Will *et al.*, 2007, 2012) is a more generally applicable method that clusters putative RNA genes according to structure as a means of improving RNAz predictions. For many genome scans, however, experiments are used to confirm a handful of the very best results, and the bulk of the remaining results are used only to estimate the number of RNA genes in the genome. Clearly, there is scope for improvement in the accuracy of the prediction of RNA genes, and it is therefore useful to evaluate the existing software on carefully defined positive and negative datasets.

I chose to evaluate RNAz (Gruber *et al.*, 2010) and EvoFold (Pedersen *et al.*, 2006) because they have different underlying methodologies, and have been used for several genome scans. The MARMOSSET pipeline provides genomic alignments of RNA genes

as a positive dataset. To create a negative dataset for the calculation of false positive rates, these genomic alignments are randomised in a way that preserves dinucleotide content and other alignment properties that are known to affect RNA gene prediction. My comprehensive assessment of these programs provides a degree of confidence in their results, demonstrates that different programs are appropriate in different contexts, and describes properties of the alignments and RNA families that affect gene prediction.

3.2 Materials and methods

3.2.1 Data

3.2.1.1 Genomic alignments of RNA genes

Three sets of RNA gene alignments were used as input to EvoFold and RNAz, derived from EPO 12-species and 35-species datasets and MultiZ 11-species datasets, as described in the previous chapter; these alignments have 400 bases of flanking on either side of the RNA gene. To evaluate the RNA gene prediction programs in a realistic manner it is important to include flanking sequence, as this tests their ability to find RNA genes in a genomic context. EvoFold and RNAz process the alignment as a series of overlapping windows, with default window sizes of 240 and 120, respectively, so a flanking length of 400 bases ensures that the programs are tested on a section of alignment without an RNA structure, but not so much that effort is wasted on analysis of bases remote from the RNA gene.

3.2.1.2 Randomised alignments

To judge the degree to which the prediction programs will erroneously predict RNA genes it is necessary to generate alignments that appear as similar as possible to the RNA alignments, but which lack the structural information of an RNA gene (Babak *et al.*, 2007; Gesell and Washietl, 2008). I use Multiperm (version 0.9.3) to generate 10 randomisations of each alignment that effectively destroy the structure of the RNA gene, but approximately preserve dinucleotide content and local conservation, and exactly preserve gap structure (Anandam *et al.*, 2009).

A problem with shuffling methods that exactly preserve dinucleotide content, which

have previously been used to test RNA gene prediction, is that many columns must remain fixed, even with pairwise alignments (Babak *et al.*, 2007). I use Multiperm to circumvent this problem, and verify that the randomisations are adequate by comparing the equivalent bases of each genomic alignment and its randomisations. The expectation is that the bases should be the same approximately 25% of the time (since I can only detect when a base has changed to a different base, not when it has “changed” to the same base), and this is indeed the case (Table 3.1).

Table 3.1: Proportion of identical bases in genomic alignments and randomisations.

	Mean	SD
EPO-12	0.22	0.05
EPO-35	0.28	0.10
MultiZ	0.25	0.04

3.2.1.3 Phylogenetic trees

EvoFold requires a tree as input so I use the neighbour-joining tree (BIONJ: Gascuel, 1997) that was calculated for each alignment, using PhyML 3.0 (Guindon *et al.*, 2010), as part of the MARMOSSET pipeline.

3.2.2 RNA gene prediction programs

3.2.2.1 EvoFold

An unpublished version of EvoFold (v.7b) was provided by the lead author of the original paper (JS Pedersen, personal communication); it uses the same model as the original version, but is faster and incorporates functionality for scoring and collating the results. (This version is not publicly available due to licensing restrictions.) The default windowing scheme, of 240 bases with an offset of 80 is used. EvoFold outputs the co-ordinates of the alignment that are considered to be RNA (or nothing if no RNA is predicted), without the need for a ‘score’ threshold to be specified.

3.2.2.2 RNAz

I use version 2.0 of RNAz, but will omit the version number hereafter, for the sake of readability. The dinucleotide null model of RNAz (rather than the default mononucleotide option) is used, as this improvement to version 2.0 of the software is expected to improve prediction accuracy; and I execute RNAz on both strands, rather than just the default forward strand. The default behaviour of RNAz selects a maximum of 6 maximally different sequences, although it can use alignments of more species. To test whether RNAz utilises the increased evolutionary information available in large alignments, RNAz is executed twice, once with the default of 6 or fewer species, and once with all available species (denoted ‘Max Sp.’ in the results). The Perl scripts that accompany RNAz are used to pre-process the alignments with default settings for the remaining parameters, including a windowing scheme of 120 bases and an offset of 40. The threshold used to identify the presence of RNA is an ‘RNA-class probability’ greater than 0.9, which is defined as a ‘high confidence’ dataset (Washietl *et al.*, 2005a).

3.2.3 Evaluation

I evaluate the programs in two ways, first recording whether an RNA gene is predicted, and if so, how well the boundary between RNA and flanking is detected. I use randomisations of the alignment to calculate the false positive rate for the case of RNA gene prediction. To evaluate boundary detection I use the flanking sequence as a negative dataset to determine the accuracy of the differentiation between RNA and flanking.

3.2.3.1 Predicting RNA genes

For each of the genomic and randomised alignments I record whether or not a prediction was made (Table 3.2). For the genomic alignments, predictions that overlap the RNA gene are counted as true positives, regardless of how much of the gene was predicted. To evaluate and compare the predictions I calculate the true positive rate (recall) and the false positive rate, and also the precision, which may be a more useful statistic than the false positive rate for researchers following up the results of RNA gene prediction (Babak *et al.*, 2007). To ensure there is no ambiguity in terminology, the statistical metrics are formally defined in Table 3.3. Confidence intervals for the true and false positive rates are calculated with the Wilson score method (Wilson, 1927; Newcombe, 1998).

Table 3.2: Predicting RNA genes.

	Genomic	Randomised
Predicted	True Positive (TP)	False Positive (FP)
Not Predicted	False Negative (FN)	True Negative (TN)

Table 3.3: Statistical definitions.

	TP Rate Recall Sensitivity	FP Rate	Precision	Specificity
Definition	$TP/(TP + FN)$	$FP/(FP + TN)$	$TP/(TP + FP)$	$TN/(TN + FP)$

3.2.3.2 Detecting RNA gene boundaries

Predicting whether an RNA gene is present somewhere in an alignment is a liberal test because it does not examine where RNA has been predicted. To provide a stricter, and potentially more informative, test of performance I analyse the boundaries of regions that are predicted to be RNA, which are provided by RNAz and EvoFold as part of their standard output. The co-ordinates of inferred genes are compared to the original alignment, site by site, to determine the number of bases that are correctly inferred to be within the known RNA. In this analysis the negative dataset is not a randomised alignment; each genomic alignment contains both positive and negative datasets, the RNA and flanking sequences respectively (Table 3.4). Note that this assumes that the sequences in the source Rfam data span the complete RNA gene.

Table 3.4: Detecting RNA gene boundaries.

	RNA	Flanking
Detected	True Positive (TP)	False Positive (FP)
Not Detected	False Negative (FN)	True Negative (TN)

3.2.4 Implementation

To automate the execution and evaluation of RNA gene prediction software, I wrote a pipeline in Perl, named **TARSIER: Testing and Analysing RNA gene Software Including Evolutionary Relationships**. The TARSIER pipeline consists of a module in the MonkeyShines library that was described in Chapter 2, and a small number of scripts in the

tarsier subdirectory of the MonkeyShines software package. The TARSIER pipeline makes it simple to apply the RNA gene prediction programs to any set of alignments, and its modular nature means that when new programs are developed, they can easily be added.

The MonkeyShines software library, incorporating the TARSIER pipeline, is freely available under an open-access licence at http://bitbucket.org/james_monkeyshines/monkeyshines/overview.

3.3 Results

3.3.1 True and false positive rates of RNA gene prediction

Figure 3.1 shows the true positive rates (TPRs) and false positive rates (FPRs), broken down by RNA gene alignments that are either within introns (or, in a few cases, immediately upstream or downstream) of protein-coding genes ('genic') and those that are not. EvoFold has low TPRs for the EPO-12 and MultiZ alignments, and a higher TPR for EPO-35 alignments, suggesting that the program needs a certain number of species before it is effective. The FPR for EvoFold tends to be higher than the TPR, so it is more likely to predict an RNA gene in a randomised alignment than an RNA gene alignment. There is little difference between the default version of RNAz with alignments limited to 6 species or those that have as many species as possible ('Max Sp. '), but since the latter tend to have slightly lower FPRs these results will be used in comparisons with EvoFold in subsequent sections. There is negligible difference between the results for RNA genes that are within or near protein-coding genes, and those that are not.

3.3.2 Detecting RNA gene boundaries

To assess how well the programs determine the boundary between RNA and flanking sequence I examine the set of predictions from each program, using sensitivity to measure how much of the prediction corresponds to the true RNA gene, and specificity to measure the amount of flanking that is excluded from the predictions (Table 3.5). Both EvoFold and RNAz predict genes in the flanking regions that do not overlap with the true RNA gene (that is, sensitivity = 0). With the exception of EvoFold and the

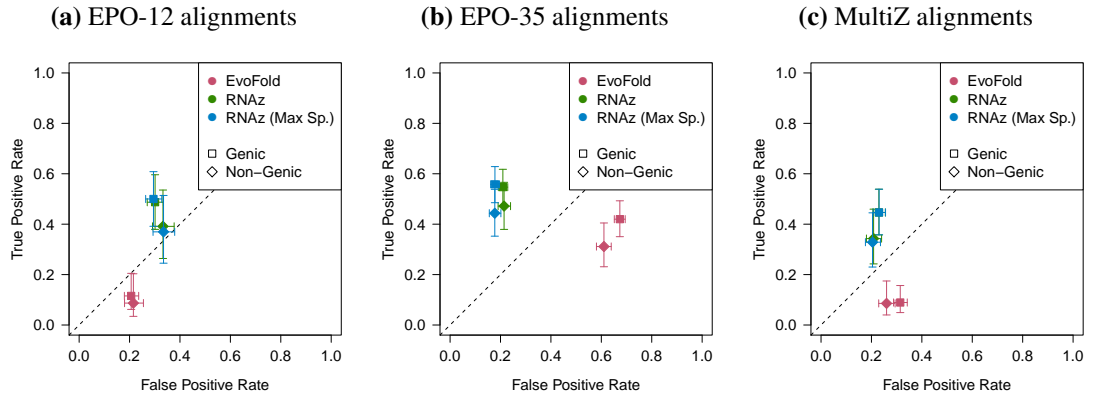


Figure 3.1: True and false positive rates of RNA gene prediction for (a) EPO-12 alignments (b) EPO-35 alignments, and (c) MultiZ alignments. Bars show the 95% confidence intervals.

EPO-35 alignments, specificity is high, meaning that boundaries are detected quite accurately when a true positive prediction is made.

Table 3.5: Mean sensitivity and specificity in detecting RNA gene boundaries, and the percentage of predictions with sensitivity = 0, where RNA is predicted in the flanking regions only. Values in parentheses are the mean sensitivity and specificity including predictions with sensitivity = 0.

		EPO-12	EPO-35	MultiZ
EvoFold	Sensitivity	0.56 (0.17)	0.80 (0.42)	0.52 (0.13)
	Specificity	0.96 (0.93)	0.39 (0.53)	0.93 (0.92)
	Sensitivity = 0	70%	47%	75%
RNAz (Max Sp.)	Sensitivity	0.84 (0.63)	0.90 (0.72)	0.92 (0.71)
	Specificity	0.87 (0.85)	0.88 (0.87)	0.89 (0.87)
	Sensitivity = 0	24%	20%	23%

To better visualise boundary detection, I calculated the prediction rate at each site for the EPO-35 alignments and their randomisations (Figure 3.2). The relatively poor specificity of EvoFold occurs because when it makes a prediction, it tends to include most of the flanking region in its prediction. In the randomised alignments, EvoFold makes more predictions in flanking regions than the RNA gene region, suggesting that the Multiperm randomisation removes structural information but that EvoFold continues to detect a signal of conservation in the flanking.

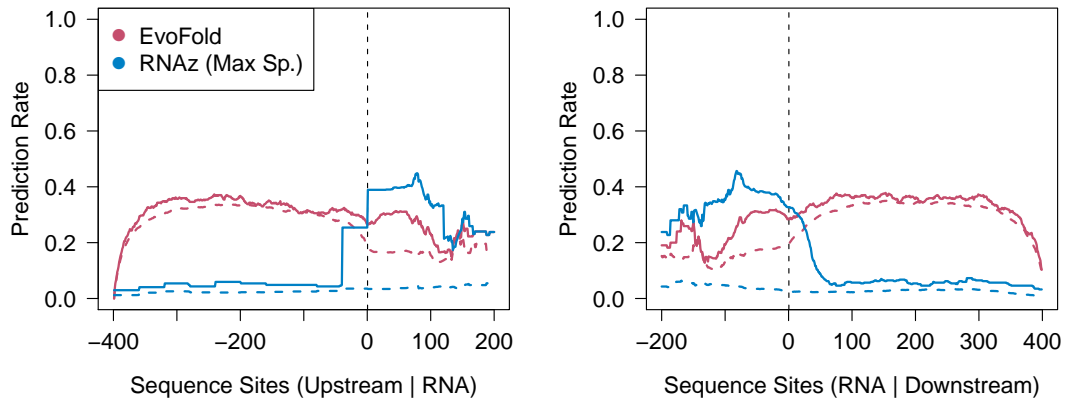


Figure 3.2: Detecting RNA gene boundaries in EPO-35 alignments. ‘Prediction Rate’ represents the number of times an RNA gene is predicted at a site, across all alignments. Solid lines represent genomic alignments, dashed lines represent randomised alignments. The vertical dotted lines indicate the boundary between the RNA gene and flanking regions.

3.3.3 Comparing prediction programs

Figure 3.3 shows the overlap between the predictions from EvoFold and RNAz, for the EPO-35 alignments. Comparisons on the EPO-12 and MultiZ alignments are not informative due to the poor performance of EvoFold on these alignments. A subset of RNA genes are correctly predicted by both programs, but they tend to have different sets of incorrect predictions. The size of the overlap between pairs of programs can be compared to the expected value under randomly distributed data, by using the hypergeometric distribution (Kim *et al.*, 2001). The overlap of true positives is significantly greater than expected by chance, by a factor of 1.26 ($p < 0.01$), and the overlap of false positives is half the expected value ($p < 0.01$).

RNAz consistently predicts RNA genes that are 120 or 160 bases in length, irrespective of whether the prediction is in the RNA region or in the flanking sequence (predictions are always a multiple of 40, due to RNAz’s windowing method). In contrast, EvoFold mostly predicts very short RNA genes in the flanking regions, but sometimes predicts RNA genes that span the whole alignment.

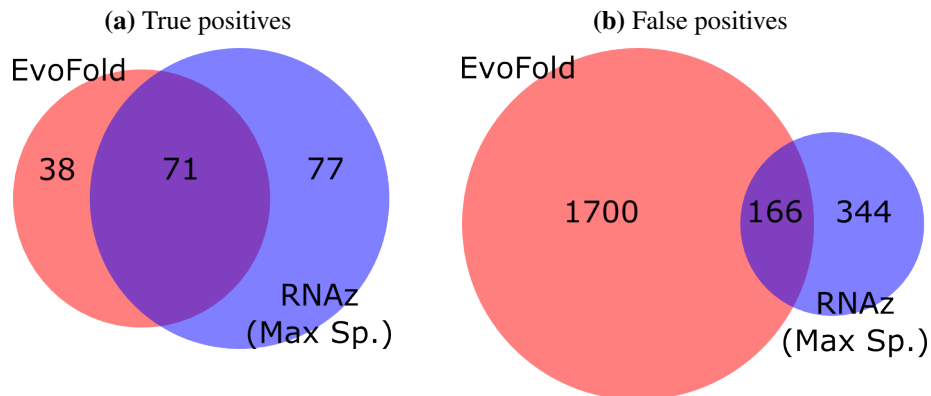


Figure 3.3: Overlap of (a) true and (b) false positive predictions between EvoFold and RNAz (Max Sp.), for the EPO-35 alignments. Images were generated with BioVenn (Hulsen *et al.*, 2008).

3.3.4 Properties that affect RNA gene prediction

The accuracy of RNA gene prediction is not affected by the GC content of an alignment (Figure 3.4a), for either RNAz or EvoFold, contrary to previous analyses (Washietl *et al.*, 2007; Torarinsson *et al.*, 2008). The percentage of paired bases in an RNA gene should affect the performance of the prediction programs, since they are designed to exploit the signal arising from secondary structure, and this is the case, with recall increasing in line with the proportion of paired bases for both prediction programs (Figure 3.4b). Precision also increases with the proportion of paired bases for RNAz, but not for EvoFold, which is a consequence of the program having a false positive rate greater than the true positive rate.

Having more species in an alignment should also be of advantage to RNA gene prediction programs, but the relationships are not straightforward. RNAz's recall is independent of the number of species in the alignment, and its precision increases with the number of species up to point, before dropping off (Figure 3.4c). In contrast, EvoFold's recall increases with species number (and precision is again uninformatively low).

3.4 Discussion

De novo prediction of RNA genes is known to be a difficult problem, and I have demonstrated the usefulness of evaluating and comparing prediction programs on real-world

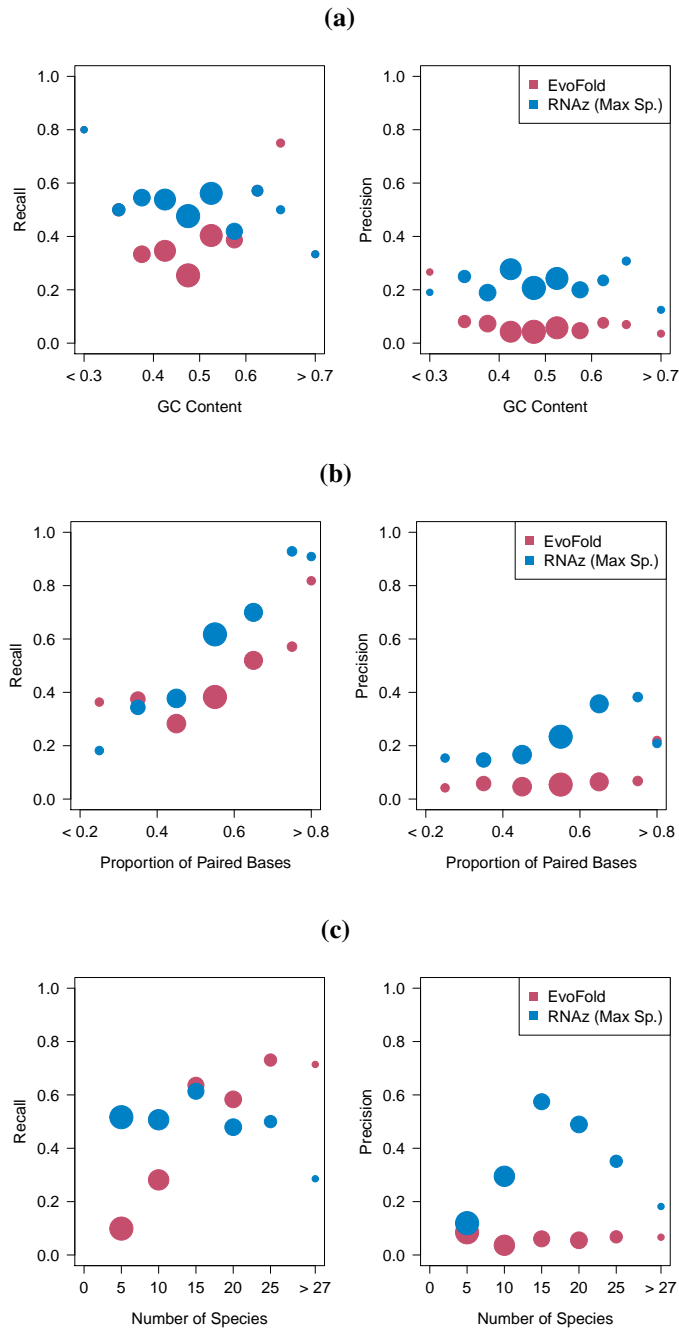


Figure 3.4: The effect of **(a)** GC content, **(b)** paired bases, and **(c)** species number on RNA gene prediction in EPO-35 alignments. ‘Recall’ is the true positive rate, and ‘precision’ is the proportion of predictions that are correct. The alignments are grouped into bins, and the area of the circles is proportional to the number of alignments in each bin.

genomic alignments of RNA genes. Before discussing the factors that affect *de novo* RNA gene prediction, it is useful to compare my results with those of similar studies. Babak *et al.* (2007) examined a range of programs including EvoFold and RNAz, but direct comparison with my results is complicated by the use of different datasets and the limitations imposed by those authors' shuffling procedure. The RNA families in the studies are similar, however, and my results qualitatively agree on the manner in which different types of RNA are differentially predicted.

In evaluations of EvoFold and RNAz on the ENCODE pilot regions (ENCODE Project Consortium, 2007), the false positive rates (FPRs) are 71% (Confidence Interval, CI: 70-72%) for EvoFold (Washietl *et al.*, 2007) and 54% (CI: 52-56%) for RNAz (Gruber *et al.*, 2010). The estimate for EvoFold does not use randomisations that preserve dinucleotide content, and thus may be an underestimate of the FPR (Babak *et al.*, 2007; Gesell and Washietl, 2008). I use a randomisation process that approximately preserves dinucleotide content, and for the EPO-35 alignments estimate an FPR of 65% (CI: 63-67%) for EvoFold, slightly lower than previous estimates, perhaps due to the strict filters that were applied to the genomic alignments. Nonetheless, a value this high is rather disappointing from an evolutionary point of view, since an explicitly phylogenetic approach has the potential to provide a greater understanding of RNA gene evolution.

For RNAz with no restrictions on the number of species, I estimate the FPR to be 31% (CI: 28-34%) for the EPO-12 alignments and 18% (CI: 16-19%) for the EPO-35 alignments, substantially lower than earlier estimates which used fewer species. The additional information on conservation in larger alignments is effectively exploited by RNAz, and this trend is also apparent, to a lesser extent, in the RNAz analysis that limited alignments to 6 species (Figure 3.1). A more sophisticated evaluation of the effect of evolution requires knowledge of how well RNA models of evolution describe the genomic alignments, which is the subject of the following two chapters.

False positive rates are difficult to interpret without accompanying estimates of the true positive rate (TPR). By using genomic alignments that contain RNA genes I am able to estimate TPRs for the prediction programs, which was not possible in previous studies that used the ENCODE regions, because these regions contain only a handful of annotated RNA genes (Washietl *et al.*, 2007). It is rather curious that a large portion of predicted RNA genes are located wholly in the flanking regions (Table 3.5). EvoFold and RNAz are able to predict multiple RNA genes in an alignment and both scan the alignment in overlapping windows of fixed size, so predictions in the flanking

region should not hinder further predictions in the RNA region. Approximately two-thirds of the RNA gene alignments are within protein-coding genes or have protein-coding genes in the flanking regions, and the predictions in the flanking regions occur with the same proportion. So it may be that the prediction programs are detecting the conservation of protein-coding regions and erroneously inferring RNA genes, but there remains an unknown source of conservation that is leading the programs astray.

How, then, should a researcher proceed if they wish to find RNA genes in some alignment? One option is to use multiple programs and to regard some intersection of the results as the most likely. The degree of overlap between the results of different programs has previously been reported to be much less than expected (Washietl *et al.*, 2007; Torarinsson *et al.*, 2008), but I find this to be the case only for the false positive predictions. The percentage of the genome that contains RNA genes is expected to be relatively small, which suggests that any genomic scan for RNA genes will be swamped by false positives, because there is greater opportunity for them to occur. Thus, the previously observed (lack of) overlap between programs could be dominated by the relatively large numbers of independently predicted false positives, masking the tendency of the programs to predict the same true positives. So, rather than choosing between prediction programs, the results of both programs can be combined, and the intersection of predictions may provide a high-confidence dataset, with a reduced false positive rate. The TARSIER pipeline that I developed can easily be used to apply EvoFold and RNAz to a set of genomic alignments, and analyse the results. The *moses* software (Raasch *et al.*, 2010) was explicitly designed to combine the results from different methods of RNA gene prediction, but it does not currently include EvoFold.

Researchers wishing to find RNA genes, or develop tools to do so, should also bear in mind the potential for poor quality alignments to produce inaccurate predictions (Saito *et al.*, 2010). Given the results of the previous chapter, that showed that the sequences and gap distributions of EPO-12 and MultiZ alignments were similar, the difference between the true and false positive rates for these alignments (Figure 3.1) is perhaps surprising. The results for EPO-12 alignments are better than the MultiZ alignments, and EPO-35 alignments are, in turn, more accurate than EPO-12, suggesting that increased quantities of genome data should enable more successful RNA gene prediction. However, it seems that evolutionary information is not being used as effectively as it might be, and the same alignments used in this chapter are analysed with phylogenetic RNA models in the next two chapters, in an effort to better understand the evolutionary processes of RNA genes.

Chapter 4

Comparing evolutionary models across state space

4.1 Introduction

It is often desirable to compare the likelihoods of different models, to evaluate their fit to the data, and thus their usefulness as descriptions of real evolutionary processes. There are a range of statistically rigorous methods to compare likelihoods within a given state space, but it is not possible to compare the likelihoods of models with different state spaces, since this effectively changes the data on which the calculations are conditioned (Burnham and Anderson, 2002). However, if there is a mapping between the states of different models, then it is possible to map all models under examination to the same state space, and thus compare them with standard statistical methods.

To compare models across state space it is necessary to define a mapping between every state in each model to at least one state in the other model. In biological terms this is usually not restrictive because the types of model that one would want to compare have natural, biologically meaningful mappings, such as that between nucleotides and codons. Such mappings were described by Whelan and Goldman (2004) to move between nucleotide and codon representations of the data, but the mappings were incidental to the main subject matter of the article, and are not explicitly considered. The mapping of nucleotide models to a 64-state space by Seo and Kishino (2009) demonstrates not only that such a mapping is possible, but that the likelihoods of these models are directly comparable. The extension of this technique to compare nucleotide and dinucleotide models (useful in the context of RNA alignments) is straightforward, and is outlined in the next chapter.

The mapping between amino acids and codons (Seo and Kishino, 2008) is an example of a relationship in which a state in one space represents a group of one or more states in another space. In this chapter I present an extension of the proof of Seo and Kishino (2008), which permits the comparison between any two models of evolution for which such a mapping is possible (and, hopefully, sensible in biological terms). This provides a foundation for selection between certain (but not all) models of evolution, such as between RY and nucleotide models, or different categories of RNA model. Using a special case of the proofs outlined below it is easy to calculate adjustments to the likelihoods to enable comparisons among physicochemical models and standard amino acid models. The reduction in the number of parameters in these physicochemical models also permits the possibility of tractable mechanistic, rather than empirical, models of amino acid substitution.

After establishing some notation and definitions, I first show that one can map from a model with a larger state space to one with a smaller state space, essentially by averaging the parameters of the first model. The main result of this chapter is the more complicated case of mapping in the other direction (note that the chapter does not have a conventional structure, since the ‘Methods’ are effectively ‘Results’). Finally, if certain parameters are restricted in the mapping then models with different state spaces can be shown to be equivalent, and a relatively simple formula can be used to generate a likelihood ‘correction’ term which is applied to make the likelihoods of the two models directly comparable.

4.2 Comparing evolutionary models across state space

4.2.1 Modelling evolution as a Markov process

Evolution can be mathematically modelled by a time-reversible Markov process that describes changes between states using a rate matrix $\mathbf{Q} = \{q_{ij}\}$, where q_{ij} is the substitution rate between states i and j . Rows in the matrix must sum to zero, which is achieved by setting $q_{ii} = -\sum_{i \neq j} q_{ij}$. The equilibrium frequency of i is denoted by π_i . The constraint of reversibility means that $\pi_i q_{ij} = \pi_j q_{ji}$, and $q_{ij} = s_{ij} \pi_j$, where $\mathbf{S} = \{s_{ij}\}$ is a symmetric matrix of exchangeability parameters; so s_{ij} describes the tendency of changes between i and j (note that $s_{ij} = s_{ji}$). To calculate L , the likelihood of a model, it is necessary to convert from a rate matrix \mathbf{Q} to a transition probability matrix over time t , $\mathbf{P}(t) = \{p_{ij}(t)\} = e^{t\mathbf{Q}}$.

4.2.2 Mapping between compound and distinct models

This chapter is concerned with mappings between ‘compound’ models and ‘distinct’ models. The states in a compound model are defined as a collection of one or more (atomic) states from an associated distinct model. Each distinct state must appear in one, and only one, of the groups that form the compound states, and at least one compound state must have more than one member. (Strictly speaking, there could be one-to-one mappings between compound and distinct states, but such a situation renders the proof redundant, and I assume fewer compound than distinct states.) Note that the definition of a distinct model in equation 4.2 below includes a term from the definition of the related compound model (equation 4.1), and thus this is not a general proof, as defining the distinct model in this way places restrictions on its parameterisation.

For brevity I will treat a compound state and the group that it represents as synonymous and refer to, for example, ‘members of the compound state’, rather than the unwieldy but more precise ‘members of the group defined by the compound state’.

4.2.2.1 Definitions

A compound model C has a rate matrix \mathbf{Q}_C , defined by:

$$q_{c_i c_j} = \begin{cases} s_{c_i c_j} \pi_{c_j} & i \neq j \\ -\sum_{k \neq i} q_{c_i c_k} & i = j \end{cases} \quad (4.1)$$

where $c_i \in C$, $1 \leq i \leq n$, are states in the model. The $m (> n)$ states of an associated distinct model are defined in relation to the states of the compound model, where membership of a compound state is indicated by the first subscript: $d_{i,1}, d_{i,2}, \dots, d_{i,l}$, $l \geq 1$, $\forall i \in \{1, 2, \dots, n\}$. The distinct model has a rate matrix \mathbf{Q}_D , defined by:

$$q_{d_{i,x} d_{j,y}} = \begin{cases} s_{c_i c_j} \pi_{d_{j,y}} & i \neq j \\ s_{d_{i,x} d_{j,y}} \pi_{d_{j,y}} & i = j \text{ and } x \neq y \\ -\sum_{k \neq i, z \neq x} q_{d_{i,x} d_{k,z}} & i = j \text{ and } x = y \end{cases} \quad (4.2)$$

where $c_i \in C$ are states in the compound model and $d_{i,x} \in D$ are states in the distinct model. Note that there are two potentially sensible simplifications to equation (4.2). If the rate of change among members of a compound state is assumed to be equal,

then the $s_{d_{i,x}d_{j,y}}$ term can be replaced with a parameter ρ_i , describing the rate of change within a group. Further, if these within-group rates of change are the same for all compound states, a single parameter, ρ , suffices.

4.2.3 Distinct models to compound models

The main result of this chapter is a proof of the mapping from a compound model to a distinct model, but it is possible to perform the mapping in the other direction, from distinct to compound models. For completeness it is useful to demonstrate the bi-directionality of the mapping between compound and distinct models, but it will not often be practicable to move from a distinct to a compound model. One would usually want to recode the input data and perform the analysis in a smaller state-space, rather than doing a more complex and time-consuming analysis in a larger state-space, and then, in essence, averaging the model parameters.

To prove that one can map from a distinct to a compound model, it must be shown that the rate matrix of the compound model can be expressed in the parameters of the associated distinct model. The formal demonstration of this mapping was given for the conversion from a codon to an amino acid model in Yang *et al.* (1998).

The rate matrix \mathbf{Q}_C of a compound model C can be expressed as follows:

$$q_{c_i c_j} = \begin{cases} \sum_{x=1}^{n_i} \sum_{y=1}^{n_j} \frac{s_{d_{i,x}d_{j,y}} \pi_{d_{i,x}} \pi_{d_{j,y}}}{\pi_{c_i} \pi_{c_j}} \pi_{c_j} & i \neq j \\ -\sum_{k \neq i} q_{c_i c_k} & i = j \end{cases} \quad (4.3)$$

where n_i and n_j are the number of distinct states in the compound states c_i and c_j , respectively. (The frequencies of the compound states are the sum of the frequencies of the relevant distinct states, $\pi_{c_i} = \sum_{x=1}^{n_i} \pi_{d_{i,x}}$, but this is not represented explicitly in equation 4.3, so as not to overburden the equation with sigmas.) The proof of equation 4.3 follows that given in Yang *et al.* (1998), generalised from amino acids and codons to any set of compound and distinct states.

Proof The substitution rate from $d_{i,x}$ to c_j represents a substitution to any of the distinct states $d_{j,y}$:

$$\begin{aligned} q_{d_{i,x}c_j} &= \sum_{y=1}^{n_j} q_{d_{i,x}d_{j,y}} \\ &= \sum_{y=1}^{n_j} s_{d_{i,x}d_{j,y}} \pi_{d_{j,y}} \end{aligned} \quad (4.4)$$

And the rate from c_j to $d_{i,x}$ is the weighted average across distinct states $d_{j,y}$:

$$\begin{aligned} q_{c_jd_{i,x}} &= \sum_{y=1}^{n_j} \frac{\pi_{d_{j,y}}}{\pi_{c_j}} q_{d_{j,y}d_{i,x}} \\ &= \sum_{y=1}^{n_j} \frac{\pi_{d_{i,x}}}{\pi_{c_j}} q_{d_{i,x}d_{j,y}} \\ &= \frac{\pi_{d_{i,x}}}{\pi_{c_j}} q_{d_{i,x}c_j} \\ &= \frac{\pi_{d_{i,x}}}{\pi_{c_j}} \sum_{y=1}^{n_j} s_{d_{i,x}d_{j,y}} \pi_{d_{j,y}} \end{aligned} \quad (4.5)$$

Applying similar logic to derive substitutions between compound states gives:

$$\begin{aligned} q_{c_jc_i} &= \sum_{x=1}^{n_i} q_{c_jd_{i,x}} \\ &= \sum_{x=1}^{n_i} \frac{\pi_{d_{i,x}}}{\pi_{c_j}} q_{d_{i,x}c_j} \\ &= \sum_{x=1}^{n_i} \frac{\pi_{d_{i,x}}}{\pi_{c_j}} \sum_{y=1}^{n_j} s_{d_{i,x}d_{j,y}} \pi_{d_{j,y}} \\ &= \sum_{x=1}^{n_i} \sum_{y=1}^{n_j} \frac{s_{d_{i,x}d_{j,y}} \pi_{d_{i,x}} \pi_{d_{j,y}}}{\pi_{c_i} \pi_{c_j}} \pi_{c_i} \end{aligned} \quad (4.6)$$

And $q_{c_i c_j}$ can be similarly defined. Let $s_{c_i c_j} = \sum_{x=1}^{n_i} \sum_{y=1}^{n_j} \frac{s_{d_{i,x}d_{j,y}} \pi_{d_{i,x}} \pi_{d_{j,y}}}{\pi_{c_i} \pi_{c_j}}$; then the rate matrix defined by $q_{c_i c_j}$ is a reversible Markov process of substitutions between compound states. ■

4.2.4 Compound models to distinct models

The transformation from an n -state compound model to an m -state distinct model is an extension of the case in Seo and Kishino (2008), in which a mapping was defined from a 20-state amino acid model to a 61-state codon model. In addition to generalising the numbers of states, I also extend Seo and Kishino's proof to include different exchangeabilities, ρ_i , for each group of distinct states rather than using a single parameter. (I think that a similar proof is possible for the case in which there is variation between distinct states in a group, but I have not been able to prove this to my satisfaction.)

For any distinct model D defined with respect to a compound model C , the transition probability matrix $\mathbf{P}_D(t)$ can be expressed in terms of the parameters of $\mathbf{P}_C(t)$ and the parameters of the rate matrices \mathbf{Q}_C and \mathbf{Q}_D :

$$p_{d_{i,x}d_{j,y}}(t) = \begin{cases} p_{c_i c_j}(t) \frac{\pi_{d_{j,y}}}{\pi_{c_j}} & i \neq j \\ p_{c_i c_j}(t) \frac{\pi_{d_{j,y}}}{\pi_{c_j}} - \frac{\pi_{d_{j,y}}}{\pi_{c_i}} e^{(q_{c_i c_j} - \rho_i \pi_{c_i})t} & i = j \text{ and } x \neq y \\ p_{c_i c_j}(t) \frac{\pi_{d_{j,y}}}{\pi_{c_j}} + \frac{\pi_{c_i} - \pi_{d_{i,x}}}{\pi_{c_i}} e^{(q_{c_i c_j} - \rho_i \pi_{c_i})t} & i = j \text{ and } x = y \end{cases} \quad (4.7)$$

Proof The compound-state rate matrix \mathbf{Q}_C (defined in Equation 4.1) can be decomposed:

$$\begin{aligned} \mathbf{Q}_C &= \mathbf{U}_C \cdot \mathbf{D}_C \cdot \mathbf{V}_C \\ &= \begin{pmatrix} \mathbf{u}_{1,1}^{(c)} & \cdots & \mathbf{u}_{1,n}^{(c)} \\ \vdots & \ddots & \vdots \\ \mathbf{u}_{n,1}^{(c)} & \cdots & \mathbf{u}_{n,n}^{(c)} \end{pmatrix} \begin{pmatrix} \lambda_1^{(c)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^{(c)} \end{pmatrix} \begin{pmatrix} \mathbf{v}_{1,1}^{(c)} & \cdots & \mathbf{v}_{1,n}^{(c)} \\ \vdots & \ddots & \vdots \\ \mathbf{v}_{n,1}^{(c)} & \cdots & \mathbf{v}_{n,n}^{(c)} \end{pmatrix} \\ &= \left(\mathbf{u}_1^{(c)}, \dots, \mathbf{u}_n^{(c)} \right) \begin{pmatrix} \lambda_1^{(c)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^{(c)} \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^{(c)T} \\ \vdots \\ \mathbf{v}_n^{(c)T} \end{pmatrix} \end{aligned} \quad (4.8)$$

where $\mathbf{u}_i^{(c)}$ and $\lambda_i^{(c)}$ are eigenvectors and eigenvalues of \mathbf{Q}_C , respectively, and $\mathbf{V}_C = \mathbf{U}_C^{-1}$, so that $\mathbf{U}_C \mathbf{V}_C = \mathbf{V}_C \mathbf{U}_C = \mathbf{I}$. The transition probability matrix for the compound states is given by the matrix exponential:

$$\begin{aligned}
\mathbf{P}_C(t) &= e^{t\mathbf{Q}_C} \\
&= \mathbf{U}_C \cdot e^{t\mathbf{D}_C} \cdot \mathbf{V}_C
\end{aligned} \tag{4.9}$$

The distinct-state rate matrix \mathbf{Q}_D (defined in Equation 4.2) is ordered in the rows (and columns) to be consistent with the ordering in \mathbf{Q}_C , such that if state c_1 is the first row in \mathbf{Q}_C , then states $d_{1,1}, d_{1,2}, \dots, d_{1,l}$ are the first rows of \mathbf{Q}_D , and so on. The initial ordering of \mathbf{Q}_C , and of the rows within a compound state in \mathbf{Q}_D is arbitrary. The distinct-state rate matrix \mathbf{Q}_D can be decomposed:

$$\begin{aligned}
\mathbf{Q}_D &= \mathbf{U}_D \cdot \mathbf{D}_D \cdot \mathbf{V}_D \\
&= \begin{pmatrix} \mathbf{u}_{1,1}^{(d_{1,1})} & \cdots & \mathbf{u}_{1,m}^{(d_{n,p})} \\ \vdots & \ddots & \vdots \\ \mathbf{u}_{m,1}^{(d_{1,1})} & \cdots & \mathbf{u}_{m,m}^{(d_{n,p})} \end{pmatrix} \begin{pmatrix} \lambda_1^{(d_{1,1})} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_m^{(d_{n,p})} \end{pmatrix} \begin{pmatrix} \mathbf{v}_{1,1}^{(d_{1,1})} & \cdots & \mathbf{v}_{1,m}^{(d_{n,p})} \\ \vdots & \ddots & \vdots \\ \mathbf{v}_{m,1}^{(d_{1,1})} & \cdots & \mathbf{v}_{m,m}^{(d_{n,p})} \end{pmatrix} \\
&= \left(\mathbf{u}_1^{(d_{1,1})}, \dots, \mathbf{u}_m^{(d_{n,p})} \right) \begin{pmatrix} \lambda_1^{(d_{1,1})} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_m^{(d_{n,p})} \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^{(d_{1,1})T} \\ \vdots \\ \mathbf{v}_m^{(d_{n,p})T} \end{pmatrix}
\end{aligned} \tag{4.10}$$

where $\mathbf{u}_j^{(d_{i,x})}$ and $\lambda_j^{(d_{i,x})}$ are eigenvectors and eigenvalues of \mathbf{Q}_D , respectively, and $\mathbf{V}_D = \mathbf{U}_D^{-1}$, so that $\mathbf{U}_D \mathbf{V}_D = \mathbf{V}_D \mathbf{U}_D = \mathbf{I}$. The eigenvalues and eigenvectors of \mathbf{Q}_D can be expressed in terms of the eigenvalues $\lambda_i^{(c)}$ and eigenvectors $\mathbf{u}_i^{(c)}$ of \mathbf{Q}_C , the parameters of \mathbf{Q}_C , and the within-group rates of change. The eigenvalues of \mathbf{Q}_D can then be expressed as:

$$\lambda_j^{(d_{i,x})} = \begin{cases} \lambda_i^{(c)} & x = 1 \\ q_{c_i c_i} - \rho_i \pi_{c_i} & x \neq 1 \end{cases} \tag{4.11}$$

The matrix \mathbf{D}_D can thus be rewritten:

$$\mathbf{D}_D = \begin{pmatrix}
\lambda_1^{(c)} & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 \\
0 & q_{c_1 c_1} & \dots & 0 & \dots & 0 & 0 & \dots & 0 \\
\vdots & -\rho_1 \pi_{c_1} & \ddots & \vdots & \dots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \dots & q_{c_1 c_1} & \dots & 0 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & -\rho_1 \pi_{c_1} & \ddots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \dots & 0 & \dots & \lambda_n^{(c)} & 0 & \dots & 0 \\
0 & 0 & \dots & 0 & \dots & 0 & q_{c_n c_n} & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots & \dots & \vdots & -\rho_n \pi_{c_n} & \ddots & \vdots \\
0 & 0 & \dots & 0 & \dots & 0 & 0 & \dots & q_{c_n c_n} \\
& & & & & & & & -\rho_n \pi_{c_n}
\end{pmatrix} \quad (4.12)$$

Proving equation 4.11, is tricky because analytical calculation of eigenvalues is often not possible for non-trivial matrices. Seo and Kishino (2008) rather sidestep the issue by stating that they “found that” the equations were true. However, the case for a 2x2 compound-state rate matrix and a 3x3 distinct-state rate matrix can be proved analytically, from which it is possible, in theory, to extend the proof to larger matrices. As this proof is rather long and incidental to the main proof, it is given in Appendix A.

It is useful to introduce a function at this point, $f(c_i)$, representing the index (i.e. row number) of $d_{i,1}$ in \mathbf{Q}_D . The eigenvectors of \mathbf{Q}_D are:

$$u_{j,k}^{(d_{i,x})} = \begin{cases} u_{i,i}^{(c)} & x = 1 \\ -\frac{\pi_{d_{i,x}}}{\pi_{d_{i,1}}} & x \neq 1, j = f(c_i) \\ 1 & x \neq 1, j = k \\ 0 & \text{otherwise} \end{cases} \quad (4.13)$$

and \mathbf{U}_D can be rewritten:

$$\mathbf{U}_D = \begin{pmatrix} u_{1,1}^{(c)} & -\frac{\pi_{d_{1,2}}}{\pi_{d_{1,1}}} & \cdots & -\frac{\pi_{d_{1,l}}}{\pi_{d_{1,1}}} & \cdots & u_{1,n}^{(c)} & 0 & \cdots & 0 \\ u_{1,1}^{(c)} & 1 & \cdots & 0 & \cdots & u_{1,n}^{(c)} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ u_{1,1}^{(c)} & 0 & \cdots & 1 & \cdots & u_{1,n}^{(c)} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ u_{n,1}^{(c)} & 0 & \cdots & 0 & \cdots & u_{n,n}^{(c)} & -\frac{\pi_{d_{n,2}}}{\pi_{d_{n,1}}} & \cdots & -\frac{\pi_{d_{n,p}}}{\pi_{d_{n,1}}} \\ u_{n,1}^{(c)} & 0 & \cdots & 0 & \cdots & u_{n,n}^{(c)} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ u_{n,1}^{(c)} & 0 & \cdots & 0 & \cdots & u_{n,n}^{(c)} & 0 & \cdots & 1 \end{pmatrix} \quad (4.14)$$

Two additional functions are helpful: $g(c_i)$ is defined to represent the index (i.e. row number) of c_i in \mathbf{Q}_C ; and $h(c_i)$ is the index of $d_{i,l}$ in \mathbf{Q}_D , where l is the number of distinct states in the compound state c_i . Then \mathbf{V}_D can be written in terms of \mathbf{V}_C and the frequency parameters:

$$v_{j,k}^{(d_{i,x})} = \begin{cases} v_{g(c_i),i}^{(c)} \frac{\pi_{d_{i,x}}}{\pi_{c_i}} & j = f(c_i) \\ 1 - \frac{\pi_{d_{i,x}}}{\pi_{c_i}} & j \neq f(c_i), j = k \\ -\frac{\pi_{d_{i,x}}}{\pi_{c_i}} & j \neq f(c_i), j \neq k, g(c_i) \leq j, k \geq h(c_i) \\ 0 & \text{otherwise} \end{cases} \quad (4.15)$$

and \mathbf{V}_D can be rewritten:

$$\mathbf{V}_D = \begin{pmatrix} v_{1,1}^{(c)} \frac{\pi_{d_{1,1}}}{\pi_{c_1}} & v_{1,1}^{(c)} \frac{\pi_{d_{1,2}}}{\pi_{c_1}} & \dots & v_{1,1}^{(c)} \frac{\pi_{d_{1,l}}}{\pi_{c_1}} & \dots & v_{1,n}^{(c)} \frac{\pi_{d_{n,1}}}{\pi_{c_n}} & v_{1,n}^{(c)} \frac{\pi_{d_{n,2}}}{\pi_{c_n}} & \dots & v_{1,n}^{(c)} \frac{\pi_{d_{n,p}}}{\pi_{c_n}} \\ -\frac{\pi_{d_{1,1}}}{\pi_{c_1}} & 1 - \frac{\pi_{d_{1,2}}}{\pi_{c_1}} & \dots & -\frac{\pi_{d_{1,l}}}{\pi_{c_1}} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \vdots & \ddots & \vdots \\ -\frac{\pi_{d_{1,1}}}{\pi_{c_1}} & -\frac{\pi_{d_{1,2}}}{\pi_{c_1}} & \dots & 1 - \frac{\pi_{d_{1,l}}}{\pi_{c_1}} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ v_{n,1}^{(c)} \frac{\pi_{d_{1,1}}}{\pi_{c_1}} & v_{n,1}^{(c)} \frac{\pi_{d_{1,2}}}{\pi_{c_1}} & \dots & v_{n,1}^{(c)} \frac{\pi_{d_{1,l}}}{\pi_{c_1}} & \dots & v_{n,n}^{(c)} \frac{\pi_{d_{n,1}}}{\pi_{c_n}} & v_{n,n}^{(c)} \frac{\pi_{d_{n,2}}}{\pi_{c_n}} & \dots & v_{n,n}^{(c)} \frac{\pi_{d_{n,p}}}{\pi_{c_n}} \\ 0 & 0 & \dots & 0 & \dots & -\frac{\pi_{d_{n,1}}}{\pi_{c_n}} & 1 - \frac{\pi_{d_{n,2}}}{\pi_{c_n}} & \dots & -\frac{\pi_{d_{n,2}}}{\pi_{c_n}} \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & -\frac{\pi_{d_{n,1}}}{\pi_{c_n}} & -\frac{\pi_{d_{n,2}}}{\pi_{c_n}} & \dots & 1 - \frac{\pi_{d_{n,2}}}{\pi_{c_n}} \end{pmatrix} \quad (4.16)$$

Equations 4.13 and 4.15 follow from eigendecomposition using equation 4.11. The transition probability matrix for the distinct states is given by the matrix exponential:

$$\begin{aligned} \mathbf{P}_D(t) &= e^{t\mathbf{Q}_D} \\ &= \mathbf{U}_D \cdot e^{t\mathbf{D}_D} \cdot \mathbf{V}_D \end{aligned} \quad (4.17)$$

into which can be substituted the matrices from equations 4.12, 4.14, and 4.16, from which it can be shown that:

$$p_{d_{i,x}d_{j,y}}(t) = \begin{cases} p_{c_i c_j}(t) \frac{\pi_{d_{j,y}}}{\pi_{c_j}} & i \neq j \\ p_{c_i c_j}(t) \frac{\pi_{d_{j,y}}}{\pi_{c_j}} - \frac{\pi_{d_{j,y}}}{\pi_{c_i}} e^{(q_{c_i c_j} - \rho_i \pi_{c_i})t} & i = j \text{ and } x \neq y \\ p_{c_i c_j}(t) \frac{\pi_{d_{j,y}}}{\pi_{c_j}} + \frac{\pi_{c_i} - \pi_{d_{i,x}}}{\pi_{c_i}} e^{(q_{c_i c_j} - \rho_i \pi_{c_i})t} & i = j \text{ and } x = y \end{cases} \quad (4.18)$$

■

4.2.5 Equivalent compound and distinct models

The proof in the previous section generates a distinct-state model that has the potential to be more informative than its associated compound-state model, as with the SK-1 model of Seo and Kishino (2008). My aim, however, is to map the compound-state model into the distinct-state space such that the likelihoods of the models can be compared directly with standard model selection techniques. The distinct-state model is equivalent to the compound-state model when the rate of change between distinct states is ‘saturated’, and the ρ_i parameters are effectively infinite. This leads to a simplification of equation 4.7:

$$p_{d_{i,x}d_{j,y}}(t) = p_{c_i c_j}(t) \frac{\pi_{d_{j,y}}}{\pi_{c_j}} \quad (4.19)$$

Equation 4.19 leads to a simple correction which can be applied to the likelihood of a compound-state model to generate the likelihood of the equivalent model in distinct-state space:

$$L_D = L_C \prod_{i=1}^T \prod_{j=1}^l \frac{\pi_{d_{ij}}}{\pi_{c_{ij}}} \quad (4.20)$$

for an alignment of T taxa and of length l , where d_{ij} is the distinct state in the i th taxon at the j th site, and c_{ij} is the associated compound state.

Proof This proof is essentially the same as that in the Appendix of Seo and Kishino (2008), modified to describe compound and distinct models with n and m states, respectively. Equation 4.20 is proved for the four taxa (distinct-state) tree shown in Figure 4.1, and can be extended to more taxa in a straightforward manner. The likelihoods at the j th site for a distinct-state model and an associated compound-state model are denoted $L_D^{(j)}$ and $L_C^{(j)}$, respectively. The compound state associated with a distinct state d is denoted c_d .

The likelihood at the j th site for the distinct-state model is then:

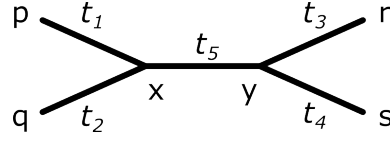


Figure 4.1: A four taxa, distinct-state, tree.

$$\begin{aligned}
L_D^{(j)} &= \sum_{x=1}^m \pi_x p_{xp}(t_1) p_{xq}(t_2) \sum_{y=1}^m p_{xy}(t_5) p_{yr}(t_3) p_{ys}(t_4) \\
&= \sum_{x=1}^m \pi_x p_{c_x c_p}(t_1) \frac{\pi_p}{\pi_{c_p}} p_{c_x c_q}(t_2) \frac{\pi_q}{\pi_{c_q}} \sum_{y=1}^m p_{c_x c_y}(t_5) \frac{\pi_y}{\pi_{c_y}} p_{c_y c_r}(t_3) \frac{\pi_r}{\pi_{c_r}} p_{c_y c_s}(t_4) \frac{\pi_s}{\pi_{c_s}} \\
&= \frac{\pi_p}{\pi_{c_p}} \frac{\pi_q}{\pi_{c_q}} \frac{\pi_r}{\pi_{c_r}} \frac{\pi_s}{\pi_{c_s}} \sum_{c_x=1}^n \pi_{c_x} p_{c_x c_p}(t_1) p_{c_x c_q}(t_2) \sum_{c_y=1}^n p_{c_x c_y}(t_5) p_{c_y c_r}(t_3) p_{c_y c_s}(t_4) \\
&= \frac{\pi_p}{\pi_{c_p}} \frac{\pi_q}{\pi_{c_q}} \frac{\pi_r}{\pi_{c_r}} \frac{\pi_s}{\pi_{c_s}} L_C^{(j)} \tag{4.21}
\end{aligned}$$

The product of all sitewise likelihoods gives Equation 4.20. ■

If the frequencies of the distinct states are unknown, each member of a compound state can be considered as equally likely. If $n_{c_{ij}}$ is the number of distinct states in the compound state c_{ij} , equation 4.20 becomes:

$$L_D = L_C \prod_{i=1}^T \prod_{j=1}^l \frac{1}{n_{c_{ij}}} \tag{4.22}$$

Note that the correction given in Equation 4.22 may be useful even if the frequencies of the distinct states are known, particularly if there are many distinct states, as it reduces the degrees of freedom in the model, which may be penalised by model selection methods. The likelihood corrections of Equations 4.20 and 4.22 are extremely useful in practice. Existing software can be used to calculate the likelihood of a compound-state model, and the only additional work to make it comparable to any other model is the definition of associated distinct states and some simple frequency calculations. In the following chapter, I use these likelihood corrections to compare a wide range of RNA models with standard model selection methodology.

As outlined in the introduction of this chapter, the proofs here can be applied to generate and compare a variety of evolutionary substitution models, but in this thesis

I am chiefly interested with RNA models. However, as an example of the applications beyond RNA models, in Appendix B I demonstrate that it is possible to directly compare a 2-state RY model with a 4-state nucleotide model, using the proofs in this chapter. It is hoped that this example also makes the workings of the proof in this chapter more understandable by instantiating the variables with some concrete values.

Chapter 5

Investigating RNA models of evolution

5.1 Introduction

Understanding the evolutionary relationships between species, genes, and populations is important in many areas of biology, and this is usually obtained through the inference of a phylogenetic tree from a set of aligned sequences. The landmark paper by Woese and Fox (1977) demonstrated that the presence of ribosomal RNA in all living organisms and its high degree of conservation make it an excellent gene for studying species relationships, and ever since it has been a popular choice for phylogenetic inference, ranging from the algae that live on sloth fur (Suutari *et al.*, 2010) to hundreds of metazoan species (Mallatt *et al.*, 2010). The biological importance of rRNA (and tRNA) is well established, but recently the significance of other types of non-coding RNA (ncRNA) has been recognised (reviewed in Griffiths-Jones, 2007; Mattick, 2009). For these genes, phylogenetic tree estimates can be used to investigate relationships within and between families of ncRNA, in order to better understand their evolution and function (e.g. Cuperus *et al.*, 2011; Wang and Ruvinsky, 2012). For example, a microRNA precursor might be subject to several, potentially antagonistic, evolutionary constraints, whereby the functional site(s) of the microRNA could be derived from one or both sides of the base-paired stem region (Berezikov, 2011).

Inferring trees from alignments of sequences requires a reliable method of inference, such as maximum likelihood or Bayesian inference (reviewed in Yang and Rannala, 2012). These methods require an explicit description of how sequences change

over time, in the form of a parametrised probabilistic substitution model. Substitution models describing nucleotide evolution typically assume that sites in an alignment evolve independently from one another, but this assumption is difficult to justify for RNA genes where there are strong functional constraints induced by complementary base-pairing in stem regions. To account for these dependencies, evolution of RNA stems is frequently described by dinucleotide substitution models, summarised by Savill *et al.* (2001). The earliest RNA models describe changes between 16 states, representing all 16 possible dinucleotides (Schöniger and von Haeseler, 1994; Muse, 1995). Later simplifications merge the 10 dinucleotides representing unstable base pairs into a single ‘mismatch’ state, resulting in models with 7 states (Tillier and Collins, 1998; Higgs, 2000). Since their inception there have been a wide variety of 16-state and 7-state RNA substitution models, each reflecting different biologically informed descriptions of RNA evolution.

In order to investigate the improvement of RNA models over their nucleotide-based counterparts, and the relative importance of their biological parameters, statistical methodology for comparing models is required. It is routine in phylogenetics for researchers to use formal model selection to decide which substitution model to use when inferring phylogenetic trees from nucleotide or amino acid sequence data (e.g. jModelTest: Posada, 2008). Common model selection methods include likelihood ratio tests for nested models and, more generally, information theoretic measures, such as AIC and BIC (Sullivan and Joyce, 2005; Burnham and Anderson, 2004). Such approaches are not appropriate for comparing models with different state-spaces, such as comparisons between 4-state nucleotide models and 7-state RNA models, or between 7-state RNA models and 16-state RNA models. When the models to be compared have a different state-spaces it changes the data on which the likelihood calculations are conditioned (Burnham and Anderson, 2002). To overcome this problem, previous studies developing RNA models have used model selection methods based on complex and time-consuming simulations (Schöniger and von Haeseler, 1999; Gibson *et al.*, 2005; Telford *et al.*, 2005), or have avoided direct model comparisons by evaluating the recovery of a ‘true’ tree by each model (Letsch and Kjer, 2011). The majority of these studies conclude that RNA models better describe the evolution of RNA stems than nucleotide models, although they usually analyse only a single alignment of rRNA (e.g. Schöniger and von Haeseler, 1994; Rzhetsky, 1995; Tillier and Collins, 1998; Savill *et al.*, 2001; Telford *et al.*, 2005; von Reumont *et al.*, 2009).

The first program to implement a wide range of RNA models was PHASE, and in

this chapter I describe some updates and modifications I made to the software, which has not been actively maintained in recent years. The first version of PHASE implements a range of 7-state RNA models and a Bayesian inference method (Jow *et al.*, 2002). The usefulness of the program is demonstrated with an analysis of stems from mitochondrial tRNA and rRNA in which the trees inferred under RNA and DNA models differ, although neither one provides a more reasonable choice than the other; model fit is not explicitly evaluated. The second version of PHASE permits the partitioning of alignments, to allow for the modelling of loops as well as stems (Hudelot *et al.*, 2003). The authors analyse the same RNA genes as in the earlier study, and report that the mean evolutionary rate for loops and stems is very similar; but the most variable loop sections are removed in a filtering step, and the unit of branch length differs (substitutions per site, as opposed to substitutions per pair), so it is hard to interpret this result. The next development of PHASE concerns mitochondrial base composition, and a three-state DNA model that groups C and T into a compound Y state (Gibson *et al.*, 2005). In evaluating whether to model separate codon positions with 3-state or 4-state DNA models, the authors test the hypothesis that the data was generated by a DNA model using a Cox test to compare against the alternative model; there is evidence to reject the DNA model for the second codon position. Leading on from this work, a non-stationary element is added to PHASE (Gowri-Shankar and Rattray, 2006, 2007), extending the modelling of rate heterogeneity with a gamma distribution to permit each rate category to have different equilibrium frequencies.

In this chapter I investigate the fit of RNA models for large numbers of mammalian RNA genes derived from genomic alignments, including many different types of ncRNA. This requires a novel method for comparing models with different state-spaces, based on methods created for comparing amino acid and codon models (Seo and Kishino, 2008, 2009), the proof of which is outlined in the previous chapter. This method enables rapid comparisons between all RNA and nucleotide models, allowing large-scale comparison without time-consuming simulation. Finally, I examine whether the choice of best-fit model affects the phylogenetic tree estimate, under the expectation that better fitting models should provide more accurate estimates. In common with previous studies, I find that RNA models very frequently provide a better fit than nucleotide models across all RNA gene alignments, with similar patterns of model fit observed for all types of ncRNA. Of the different types of RNA model, the models that describe general base pair stability, rather than the precise identity of base pairs, tend to provide a better fit than other RNA models. The choice of model can have

a substantial effect on the tree estimate, with the greatest differences being between nucleotide and RNA models, but there is also substantial variation within the different types of RNA model.

5.2 Materials and methods

5.2.1 Substitution models

5.2.1.1 Definitions

In all of the models used in this chapter, changes between states are described by a time-reversible Markov process, with rate matrix $\mathbf{Q} = \{q_{ij}\}$, where q_{ij} is the substitution rate between states i and j (reviewed in Yang, 2006). The equilibrium frequency of states is denoted by $\pi = \{\pi_i\}$, where π_i is the frequency of state i . The constraint of reversibility enforces $\pi_i q_{ij} = \pi_j q_{ji}$, and allows \mathbf{Q} to be represented as $q_{ij} = s_{ij} \pi_j$, where $\mathbf{S} = \{s_{ij}\}$ is a symmetric matrix of exchangeability parameters ($s_{ij} = s_{ji}$), which describes the relative rate of change between i and j . To calculate the likelihood of a model, L , requires the creation of a transition matrix from the instantaneous rate matrix by $\mathbf{P}(t) = \{p_{ij}(t)\} = e^{\mathbf{Q}t}$, which describes the probability of change between states i and j over a branch of length t . I use numerical superscripts to denote the dimension of a matrix and any values derived from that matrix; for example $\mathbf{Q}^4 = \{q_{ij}^4\}$ denotes a 4-state instantaneous rate matrix.

5.2.1.2 Nucleotide and dinucleotide models

This study uses eighteen different parametrisations of \mathbf{Q} to define ‘foundation models’ of nucleotide and dinucleotide evolution, which are later combined to provide a range of substitution models describing RNA evolution. To describe the evolution of independent nucleotides I use two common 4-state foundation models: the HKY model (Hasegawa *et al.*, 1985), and the general time-reversible (GTR) model (Lanave *et al.*, 1984; Tavaré, 1986). Both nucleotide foundation models are always used in conjunction with Γ -distributed rates-across-sites, indicated by a ‘+ Γ ’ suffix (Yang, 1994b). To describe evolution in base pairs I examine a range of foundation models over two different state-spaces: 16-state foundation models describing substitutions between all possible base pairs, and 7-state foundation models describing substitutions between the six stable canonical base pairs (A:U, C:G, and G:U) and a mismatch state, which contains the ten other base pairs (A:C, A:G, C:U, A:A, C:C, G:G, and U:U). Following

the naming convention of Savill *et al.* (2001), I investigate nine 16-state dinucleotide foundation models (16A, 16B, 16C, 16D, 16E, 16F, 16I, 16J, and 16K) and seven 7-state dinucleotide foundation models (7A, 7B, 7C, 7D, 7E, 7F, and 7G). The original authorship of these models is provided by Savill *et al.* (2001), with the exception of 7G, which I propose here as a natural simplification of 7E and 7F. Under 7G the instantaneous rate matrix is defined as:

$$\mathbf{Q} = \begin{pmatrix} & AU & GU & GC & UA & UG & CG & MM \\ AU & * & \pi_{G:U} & 0 & 0 & 0 & 0 & \pi_{MM}\alpha \\ GU & \pi_{A:U} & * & \pi_{G:C} & 0 & 0 & 0 & \pi_{MM}\alpha \\ GC & 0 & \pi_{G:U} & * & 0 & 0 & 0 & \pi_{MM}\alpha \\ UA & 0 & 0 & 0 & * & \pi_{G:U} & 0 & \pi_{MM}\alpha \\ UG & 0 & 0 & 0 & \pi_{A:U} & * & \pi_{G:C} & \pi_{MM}\alpha \\ CG & 0 & 0 & 0 & 0 & \pi_{G:U} & * & \pi_{MM}\alpha \\ MM & \pi_{A:U}\alpha & \pi_{G:U}\alpha & \pi_{G:C}\alpha & \pi_{A:U}\alpha & \pi_{G:U}\alpha & \pi_{G:C}\alpha & * \end{pmatrix} \quad (5.1)$$

where $\pi_{A:U} = \frac{\pi_{AU} + \pi_{UA}}{2}$, $\pi_{G:U} = \frac{\pi_{GU} + \pi_{UG}}{2}$, $\pi_{G:C} = \frac{\pi_{GC} + \pi_{CG}}{2}$, and π_{MM} is the total frequency of the mismatch states. Note that I do not examine the early 6-state models, such as those proposed by Tillier and Collins (1995), because it seems unreasonable to recode unstable base pairs as missing data, rather than explicitly incorporate them into the model.

Figure 5.1 summarises the parameterisation of the 18 foundation models described above, and how they can be grouped into four classes depending on how they deal with paired bases. The first class (red), consisting of HKY+ Γ and GTR+ Γ , ignores base-pairing and allows nucleotides to evolve independently. The remaining three classes are determined by how they describe the selective pressures acting on dinucleotides, primarily defined by the parametrisation of frequencies. The foundation models contained in the ‘All Pairs’ class (purple) consider changes between the 16 possible dinucleotides, allowing each dinucleotide, XY, to have its own equilibrium frequency, π_{XY} .

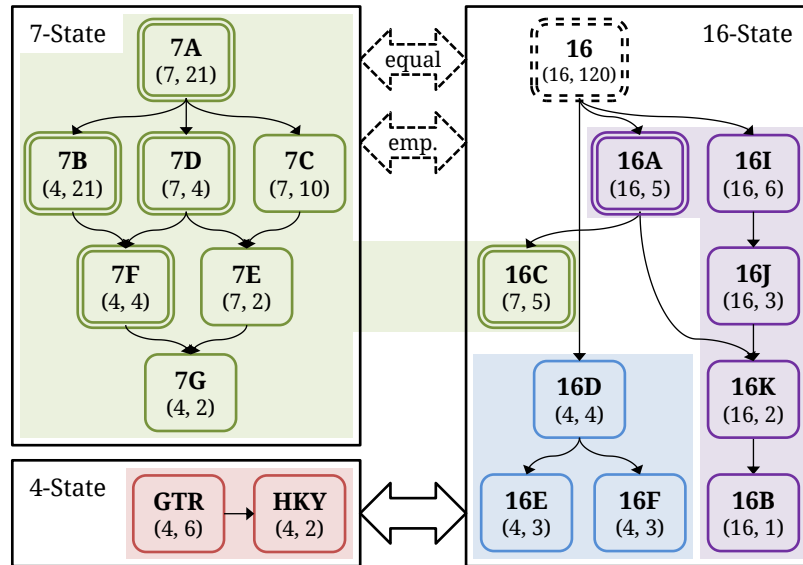


Figure 5.1: Summary of the parameterisation of 7-state and 16-state RNA and 4-state DNA models, and the relationships between them. The values below each model name are the number of frequency and exchangeability parameters, respectively. Double borders around models indicate that double substitutions are permitted. Arrows between models indicate nesting. Colouring indicates the groups of models defined in the text: DNA models (red); *All Pairs* (purple); *Stable Pairs* (green); *Stable Sets* (blue). The general 16-state model (dotted box) has too many parameters to be tractable, and is not included in this analysis. The 4-state and 16-state models are directly comparable. The 7-state models require a likelihood adjustment value to account for the mapping from 1 mismatch state to 10, which can use either equal frequencies (0 degrees of freedom) or empirical frequencies (9 d.f.).

The ‘*Stable Pairs*’ class has models with separate frequencies for each of the stable base pairs (π_{AU} , π_{UA} , π_{CG} , π_{GC} , π_{GU} , π_{UG}) and groups the 10 mismatch base pairs together into a single frequency parameter (π_{MM}). This restriction is simple in 7-state dinucleotide models because each state has its own frequency, whereas dinucleotide frequencies for the 10 mismatch states in 16C are defined as $\pi_{MM}/10$. Note that models 7B, 7F and 7G place the further restriction of strand symmetry, resulting in three frequencies for the stable base pairs ($\pi_{AU} = \pi_{UA}$, $\pi_{CG} = \pi_{GC}$, and $\pi_{GU} = \pi_{UG}$) and a single frequency describing mismatches (π_{MM}).

Finally, the ‘*Stable Sets*’ foundation models (blue) define their equilibrium frequencies based on the product of the individual nucleotide frequencies and two parameters describing the tendency for stable base pairs to occur (λ) and for wobble pairings to occur (ϕ). The formulation of these models means that the instantaneous rate of change between dinucleotides for the *Stable Sets* is different to the other two classes (See Savill

et al., 2001, for full details of all dinucleotide models). The instantaneous rate matrices for all dinucleotide models are given in Appendix C. As an alternative to mathematical formulations, Figure 5.2 shows graphical representations of some example substitution models with different numbers of states.

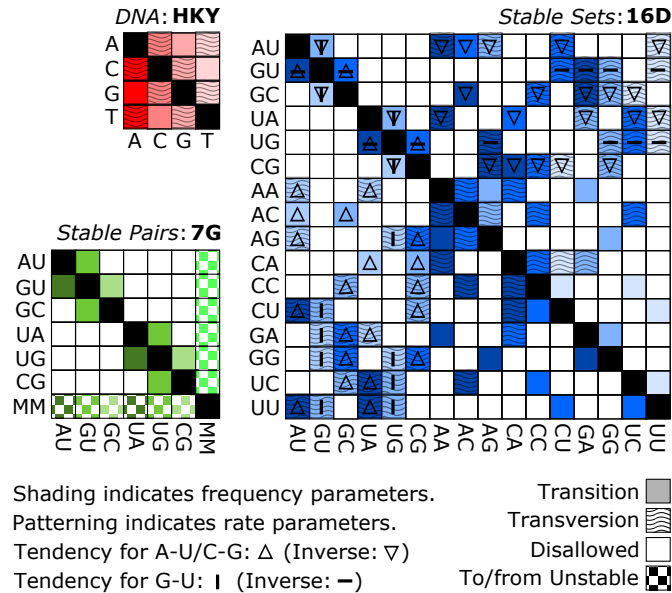


Figure 5.2: Graphical representations of example substitution models with 4, 7, and 16 states.

5.2.1.3 Modelling RNA evolution

The foundation models described above are combined to create RNA substitution models. The loop regions of the RNA are specified by the secondary structure associated with each alignment and may be modelled by either of the two nucleotide foundation models (HKY+ Γ or GTR+ Γ). The base-paired stems may be modelled by either of the 2 nucleotide foundation models, or by any of the 16 dinucleotide foundation models with or without Γ -distributed rates-across-sites, yielding $(2 + (2 \times 16) =)$ 34 possible stem models. The different combinations of stem and loop foundation models produces $(2 \times 34 =)$ 68 mixed models. A further two, non-mixed, models are also used, in which a single nucleotide model (HKY+ Γ or GTR+ Γ) is used, ignoring the loop and stem partitions. For models where the loops and stems are partitioned, a scaling factor, μ describes the evolutionary rate of stems relative to that of loops. This scaling factor can then be used to calculate meaningful tree lengths in terms of expected number of substitutions per nucleotide from RNA models combining nucleotide and dinucleotide foundation models, $\text{tree length} = (Pr(\text{loop}) \times \text{nucleotide tree length}) + 2(Pr(\text{stem})) \times$

$\mu \times$ nucleotide tree length, where $Pr(\text{loop})$ and $Pr(\text{stem})$ are the relative proportions of the nucleotides in loops and stems, respectively, and the ‘2’ corrects for dinucleotide models being scaled to evolve at one substitution per dinucleotide per unit time.

5.2.2 Model comparison

To compare the different RNA substitution models I use the corrected version of Akaike’s Information Criterion (AICc: Akaike, 1974; Burnham and Anderson, 2002). An approximation to the sample size is computed by counting the characters in an alignment, treating each base pair as a single character in the case of RNA models, following the approach of Posada and Buckley (2004). Since it is not valid to compare likelihoods computed in different state-spaces, AICc values cannot be compared between the groups of 4-state DNA models, 7-state RNA models, and 16-state RNA models (Burnham and Anderson, 2002). Previous research has used sophisticated simulation schemes to compare models (e.g. Savill *et al.*, 2001; Telford *et al.*, 2005). Instead, I project 4-state and 7-state models to a 16-state space (Figure 5.3), which then permits valid likelihood comparisons. This technique has been previously described for transforming DNA, amino acid, and codon models into 64-state models (Whelan and Goldman, 2004; Seo and Kishino, 2008, 2009). In Chapter 4 I described the mapping between 7-state and 16-state models, and in the following section I demonstrate that any 4-state model is comparable to any 16-state model.

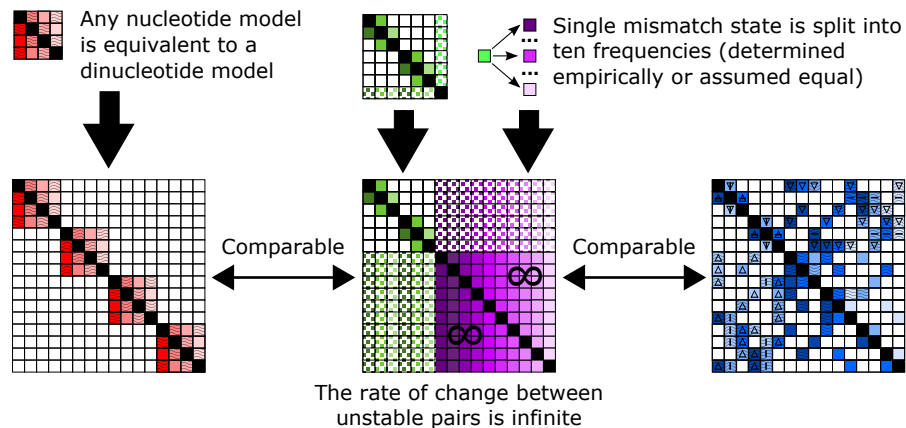


Figure 5.3: Schematic of mapping from 4-state and 7-state models to 16-state space.

5.2.2.1 Comparing 4-state and 16-state models

Previous studies have shown that 4-state nucleotide models and 64-state codon models are directly comparable (Whelan and Goldman, 2004; Seo and Kishino, 2009). In order to show that 4-state nucleotide and 16-state dinucleotide models are directly comparable I follow closely the proof of Seo and Kishino (2009). A dinucleotide model in which one nucleotide is fixed is equivalent to a 4-state model for the unfixed nucleotide:

$$q_{ij}^{16} = \begin{cases} 0 & i_1 \neq j_1, i_2 \neq j_2 \\ q_{i_2 j_2}^4 & i_1 = j_1, i_2 \neq j_2 \\ q_{i_1 j_1}^4 & i_1 \neq j_1, i_2 = j_2 \end{cases} \quad (5.2)$$

where i, j are dinucleotides, and i_1 and i_2 are the nucleotides at the first and second position of the dinucleotide, respectively. The diagonal entries of the matrix are defined by the constraint that the row sum is 0. The matrix \mathbf{Q}^{16} derived from formula (5.2) can be decomposed into two matrices, $\mathbf{Q}^{16,1}$ and $\mathbf{Q}^{16,2}$, which describe the transition rates of the first and second nucleotide, respectively. These two matrices are commutative, so $\mathbf{P}^{16}(t) = e^{t\mathbf{Q}^{16}} = e^{t(\mathbf{Q}^{16,1} + \mathbf{Q}^{16,2})} = e^{t\mathbf{Q}^{16,1}} e^{t\mathbf{Q}^{16,2}} = \mathbf{P}^{16,1}(t) \mathbf{P}^{16,2}(t)$.

The rows (or columns) of any \mathbf{Q} matrix can be interchanged without affecting the validity of the matrix, allowing the rearrangement of the rows and columns of $\mathbf{Q}^{16,x}$ ($x \in \{1, 2\}$) to obtain ‘diagonal block’ matrices which have \mathbf{Q}^4 on the diagonal and zeroes elsewhere (left panel of Figure 5.3). The calculation of $e^{t\mathbf{Q}^{16,x}}$ is then equivalent to a diagonal block matrix with $\mathbf{P}^4(t)$ on the diagonals, and the rows and columns of $\mathbf{P}^{16,x}(t)$ can subsequently be rearranged to restore their original order. Finally, multiplying $\mathbf{P}^{16,1}(t)$ and $\mathbf{P}^{16,2}(t)$ gives the original matrix $\mathbf{P}^{16}(t)$ leading to:

$$p_{ij}^{16}(t) = p_{i_1 j_1}^4(t) p_{i_2 j_2}^4(t) \quad (5.3)$$

Following the proof of equation 11 in the appendix of Seo and Kishino (2009), it is possible to derive $L^4 = L^{16}$ using my equation 5.3, and demonstrate that the likelihoods of 4-state and 16-state models are directly comparable.

5.2.2.2 Comparing 7-state and 16-state models

The likelihoods of 7-state and 16-state models cannot be directly compared, but it is relatively easy to calculate a likelihood correction value that corresponds to projecting

the 7-state model to 16-state space. Following the proof of a mapping between models with different state spaces in the previous chapter, I define the off-diagonal values of a 16-state matrix in terms of parameters from a 7-state matrix:

$$q_{ij}^{16} = \begin{cases} s_{ij}^7 \pi_j & i \in C, j \in C \\ s_{im}^7 \pi_j & i \in C, j \notin C; \text{ or } i \notin C, j \in C \\ \rho \pi_j & i \notin C, j \notin C \end{cases} \quad (5.4)$$

where i, j are dinucleotides, C is the set of canonical dinucleotides, and m is the compound mismatch state in the 7-state model. The substitution rate between mismatches is undefined in the 7-state model, so in the 16-state model I define it in terms of the dinucleotide frequency, π_j and a new exchangeability parameter, ρ , which describes the rate that mismatch dinucleotides substitute one another.

Following the work of Seo and Kishino (2008), it is possible to optimise ρ , which would create a new class of RNA models that lie somewhere between 7-state and 16-state models. I do not investigate this possibility here, however, because the rate of change between mismatches is of limited interest and including it would introduce a large number of additional models to this analysis. Instead, I concentrate on making existing 7-state models comparable with 16-state models. I assume that ρ in equation (5.4) is infinite, so that the substitution rates between all mismatch states are identical, giving the following likelihood correction:

$$L^{16} = L^7 \prod \prod^{taxa\ length} \frac{\pi_i}{\pi_m} \quad (5.5)$$

where $\pi_i, i \notin C$ is the frequency of a specific mismatch dinucleotide, and π_m is the combined frequency of all mismatch dinucleotides. Projecting the single mismatch state of the 7-state models into ten distinct states means that each of the frequencies needs to be defined. I investigate this projection in two ways, by using empirical frequencies, and by assuming that all non-canonical dinucleotides are equally likely, so that $\pi_i = \pi_m/10$. The former method introduces 9 additional parameters for the AICc calculations.

5.2.3 Implementation

For model comparisons I use a maximum likelihood (ML) approach on a fixed tree topology, with branch lengths and model parameters estimated from the data. To conduct tree search I use Bayesian MCMC analysis to obtain samples from the posterior distribution across all parameters, including trees, branch lengths, and model parameters. The results from the ML inference are used as the starting point for the MCMC, followed by 150,000 burn-in iterations. In total I perform 300,000 sampling iterations, with a sampling period of 100, yielding 3000 posterior samples. Under ML and Bayesian inference, the (di)nucleotide frequency estimates are obtained from empirical counts from the sequence data, with no subsequent optimisation.

5.2.3.1 Modifications to the PHASE software

Phylogenetic analyses are performed with a modified version of the PHASE 2.0 software package (Hudelot *et al.*, 2003; Telford *et al.*, 2005; Gowri-Shankar and Rattray, 2007), which I will refer to as version 2.1, although note that this is not (yet) a version that has been sanctioned by the original authors. PHASE 2.1 consists of improvements to the stability of the programs when using relatively short alignments and/or highly parametrised models, and some new functionality in the maximum likelihood programs (*mlphase* and *optimizer*).

The proof in Chapter 4 requires the use of empirical (di)nucleotide frequencies, but PHASE 2.0 lacks this functionality for ML inference, so I implemented this as an optional calculation. With short alignments or with models with many parameters, PHASE 2.0 can perform unreliably and crash, due to problems with the optimisation which are resolved in version 2.1. This version also automatically calculates the likelihood correction described in the previous section, to enable quick and easy comparison between the models across state space. Model selection is further simplified by the inclusion in PHASE 2.1 of a Perl script that automatically applies a range of models and calculates AIC values, given an alignment, a structure and a tree. The changes in PHASE 2.1 are detailed in Appendix D.

This unofficial version 2.1 of Phase is freely available at <http://www.monkeyshines.co.uk/phase>.

5.2.4 Genomic alignments of RNA genes

The flanking regions are removed from the genomic alignments generated in Chapter 2, and are analysed with each RNA model. For each sequence in the alignments, if one base from a pair (as indicated by the associated secondary structure) is a gap character, then I change the other base to be a gap too, and if this introduces columns that consist entirely of gap characters then these are removed. The model selection analysis requires a fixed phylogenetic tree, and I use the neighbour-joining tree generated in Chapter 2, under the assumption that this a reasonable estimate of the evolutionary relationships between the sequences.

5.3 Results

5.3.1 RNA models describe evolution better than DNA models

The results for the three different alignment sets, EPO-12, EPO-35 and MultiZ, show the same patterns, so I will focus on the EPO-35 results, since this is the largest dataset. Table 5.1 shows the best fitting model for the 287 RNA gene alignments in the EPO-35 dataset. Almost all alignments (98%) are best described by an RNA model that explicitly describes dinucleotide evolution in the stem region. Two models best describe evolution in half of the alignments, the simplest *Stable Pairs* model, 7G, and the most complex *Stable Sets* model, 16D. The 7G model is, in fact, the simplest of all the RNA models (Figure 5.1), with only 4 free parameters (equation 5.1), and tends to be selected in the most conserved alignments. The rarely selected HKY model has the same number of parameters as 7G, and this suggests that even when there are relatively few changes observed in an alignment, an RNA model provides a better description than a DNA model.

Table 5.1: Number of models with $\Delta\text{AICc} = 0$.

<i>Model Class</i>	Loop Model		
	Stem Model	HKY+ Γ	GTR+ Γ
<i>DNA</i>			
One DNA model	6	0	6
Two DNA models	0	0	0
<i>Stable Pairs</i>			
16C	18	5	23
7C	4	0	4
7E	7	1	8
7F	1	0	1
7G	58	9	67
<i>Stable Sets</i>			
16D	93	27	120
16E	33	8	41
16F	12	2	14
<i>All Pairs</i>	2	1	3
Total	234	53	287

In the few cases where a DNA model is selected, it is always a single model covering loop and stem, rather than a model partitioned for stems and loops. In the 281 alignments where an RNA model is chosen, the loop regions are best described by the simpler HKY+ Γ , rather than GTR+ Γ , in 234 (83%) alignments. The best-fit RNA models rarely include rates-across-sites heterogeneity, with only 14% of alignments using ‘+ Γ ’ dinucleotide models, suggesting that all base pairs in a stem tend to evolve at a similar rate. This observation notably contrasts with the tendency for nucleotide (Arbiza *et al.*, 2011) and amino acid (Goldman and Whelan, 2002) alignments to require spatial rate heterogeneity.

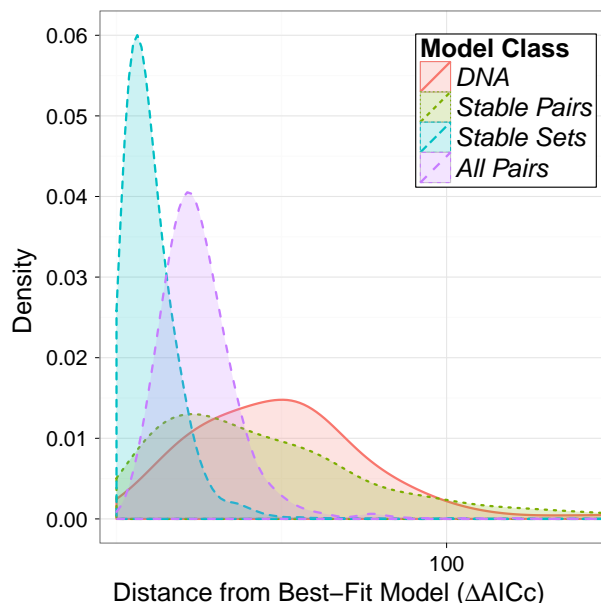


Figure 5.4: Distribution of AICc values relative to the best fit model (ΔAICc). The best fit model is not included in the plot. The x-axis is truncated at 150 for clarity.

Simply examining the best-fit model may be misleading, because when there are several similarly fitting models small differences in the likelihood may lead to different models being chosen. Figure 5.4 shows the distribution of AICc values for each class relative to the best model. In cases where the *Stable Sets* models are not selected as the best models, their AICc values tend to be very close to the best fitting model, suggesting that they consistently provide a good fit to the data even if they are not the absolute best model. The *Stable Pairs* class is much more inconsistent; in some cases it fits well, but in others it fits very poorly. Although 7G is often chosen as the best fit model, in the remaining cases it does not fit as well as the *Stable Sets* models, especially 16D, which is the first or second choice model for 242 (85%) RNA gene alignments.

The parameter estimates obtained from the dinucleotide foundation models provide some insight into RNA function and evolution. The empirical frequency of Watson-Crick base pairs is 80%, with the remaining base pairs consisting primarily of wobble base pairs (13%) and a smaller proportion of mismatches (7%). These frequency estimates are used directly by the *Stable Pairs* models, and the strong preference for 16D over other models demonstrates the overwhelming importance of differentiating Watson-Crick and wobble base pairs from each other and the mismatch base pairs, via the λ and ϕ parameters.

The frequency estimates and the best-fit models both demonstrate, as expected, that there is consistent and strong evidence for stable stems, and that wobble pair-pairing is a viable intermediate during RNA evolution. Although mismatches do occur, albeit relatively infrequently, the very low frequency (1%) with which *All Pairs* models are chosen suggests that the exact identity of mismatches when they occur is unimportant. Examining the relative rate of per nucleotide substitution in loops and stems, just under half of the RNA genes (49%) have a faster rate in stems than in loops; in many cases the difference is small, but 21% of the RNA genes have a stem rate over twice that of the loop rate.

5.3.2 Factors determining model choice

The type of RNA gene has some affect on model choice (Table 5.2), but in cases where there is more than one example of an RNA type, no single class of models is exclusively chosen. Rather than having a direct relationship with the type of RNA gene, model choice is related to the amount of structural and evolutionary information available. In the few cases where they are selected, the DNA models mostly describe evolution in snoRNA that have relatively few base pairs.

Table 5.2: Best-fit models for EPO-35 alignments, classified by RNA type. The ‘Other’ type is a heterogeneous mixture of molecules such as cis-regulatory elements and selenocysteine insertion sequences that do not naturally fit into other groups.

RNA Type	Model Class			
	DNA	Stable Pairs	Stable Sets	All Pairs
Long ncRNA	0	9	15	0
microRNA	0	33	71	2
Ribosomal	0	0	1	0
RNase P	0	1	0	0
scaRNA	0	2	9	0
snoRNA	4	52	62	1
Spliceosomal	0	1	0	0
tRNA	0	1	0	0
Vault	0	0	1	0
Other	2	4	16	0

The *Stable Pairs* models are selected most often when fewer evolutionary events are detectable in the data; as greater numbers of substitutions are inferred, on larger numbers of paired bases, the *Stable Sets* models tend to dominate (Figure 5.5). Factors such as GC content (Figure 5.5) and the number of gaps in an alignment (data not shown) do not lead to a preference for one category of model over another.

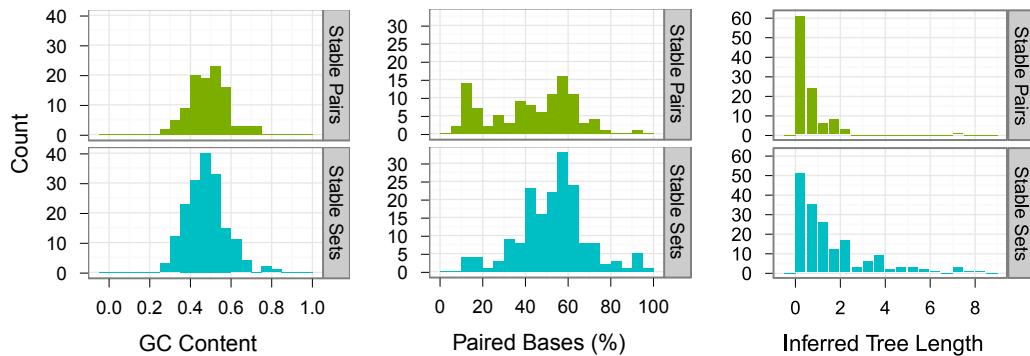


Figure 5.5: Factors affecting model choice, for the *Stable Pairs* and *Stable Sets* models: GC content; percentage of paired bases; and inferred tree length, where tree length is the sum of the individual branch lengths under the best-fit model.

5.3.3 Model choice affects tree inference

To study the effect of model choice on tree inference, I use Bayesian inference to estimate a tree under each model, for each of the RNA gene alignments. An MCMC sampling approach recovers a set of 3000 sampled tree topologies for each model and alignment. The mean overlap of sampled trees from different models indicates the similarity between the trees inferred by different models, and varies substantially between models (Figure 5.6). Some models, such as the more parameter-rich 7-state models, tend to produce trees which are different from all of the other models.

Evaluating the overlap of sampled trees demonstrates that tree estimates are different, but does not indicate how different the trees are. To measure absolute topological difference I calculated the mean Robinson-Foulds (RF) distance (Robinson and Foulds, 1981) between sampled trees, averaged across all of the alignments and normalised to account for different numbers of taxa. The results from this analysis mirror those from the overlap (Figure 5.6), with greater distances for those models with the lowest mean overlap.

It is surprising that the inferred tree is often quite different from the species tree that is distributed with the EPO-35 alignments, so I use the AU-test (Shimodaira, 2002), a less conservative version of the SH-test (Shimodaira and Hasegawa, 1999), to evaluate

whether the trees are significantly different. For 124 (43%) of the EPO-35 alignments the consensus tree from the Bayesian analysis, for the best-fit model from the ML analysis, is significantly different from the species tree. This result could indicate that the RNA gene alignments contain paralogs rather than orthologs, despite my efforts to avoid them in Chapter 2.

In the cases where RNA genes are within protein-coding genes, it is useful to examine the protein-coding regions for evidence of paralogs. I retrieve the nucleotide sequences of the protein coding regions from the EPO-35 alignments with the Ensembl Perl API, then infer trees under a GTR+ Γ model, using PhyML 3.0 (Guindon *et al.*, 2010). Of the 124 RNA gene alignments in which inferred and species trees are significantly different, 65 are within protein-coding genes, and in 11 of these cases the inferred protein-coding tree is also significantly different from the species tree, which could be due to either the presence of paralogs or misalignment. This is encouraging in that the RNA gene alignments show relatively little evidence of paralogy or misalignment, but the issue of significantly different trees remains unresolved.

The occurrence of significantly different trees does not correlate with any alignment properties, such as conservation, RNA type, or RNA length, and in most of these cases (73%), trees inferred under a DNA model are also significantly different from the species tree. Since it is unlikely that so many RNA genes have evolutionary histories different from the species tree, it may be that existing RNA models, while generally better than DNA models, still fail to correctly interpret some signals of evolution in RNA genes (Letsch *et al.*, 2010).

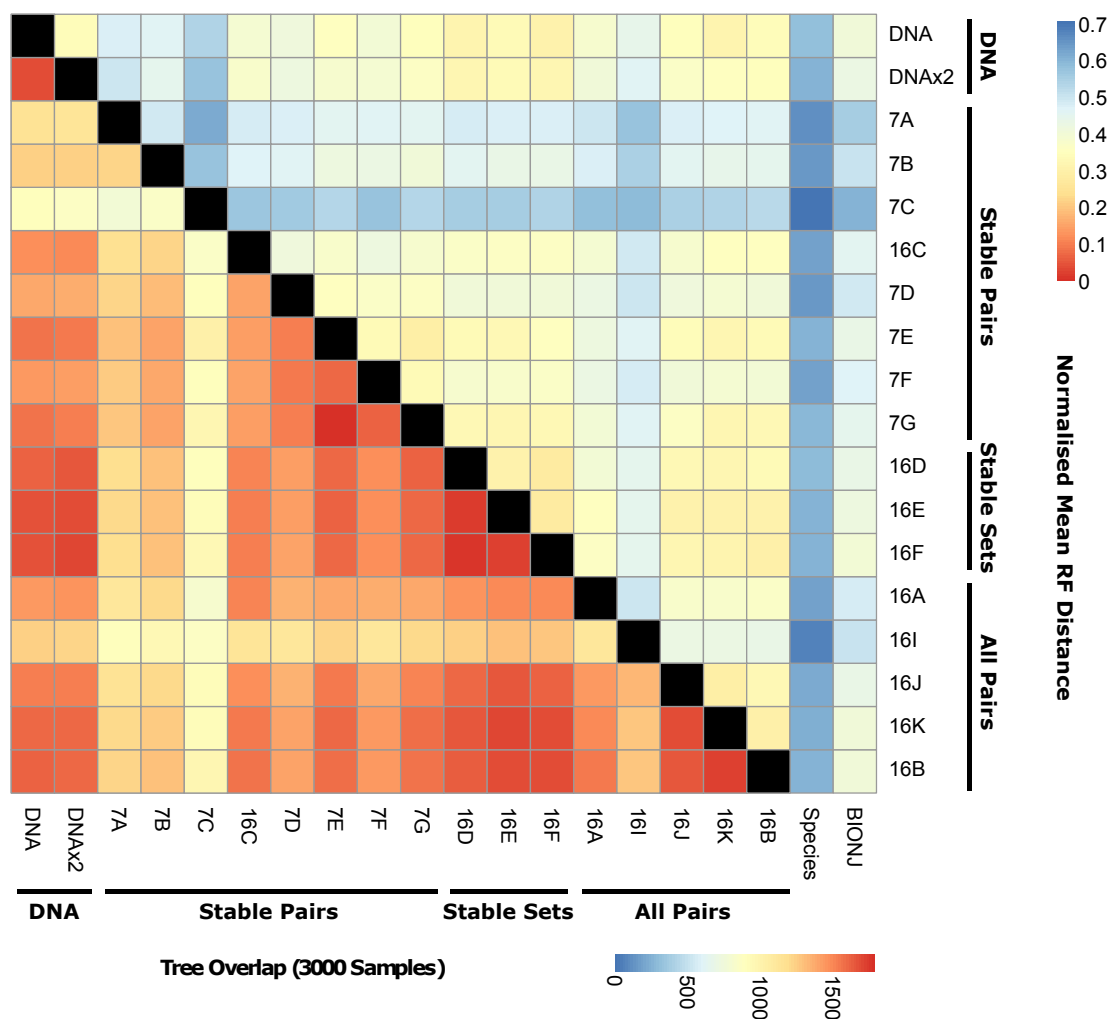


Figure 5.6: Effect of model choice on tree inference. Rows and columns in the heatmap represent different models. Data is shown for HKY loop models and ‘- Γ ’ dinucleotide models. Within a class, the models are listed in an order that approximates decreasing model complexity, from left to right (top to bottom). The lower-left triangle shows the mean overlap between the 3000 sampled trees from each MCMC tree search. The upper-right triangle shows the mean Robinson-Foulds (RF) distance between sampled trees, normalised by the number of branches in the tree so that alignments with different numbers of taxa are comparable. The trees are compared to the EPO-35 species tree, and a neighbour-joining BIONJ tree.

5.4 Discussion

A range of RNA models have been proposed by others (collated in Savill *et al.*, 2001), and in this chapter I propose a further model at the bottom of the hierarchy of 7-state models, 7G, as a natural combination of two existing models. The 7G model

is selected as the best-fit for many RNA gene alignments, and the small number of parameters in the model seem suited to describe relatively well-conserved alignments. It is noteworthy that the HKY model has the same number of parameters, but is almost never selected as the best-fit model, indicating that even when there are few observable changes, models that describe RNA-specific evolutionary processes are useful.

I define a classification of RNA models into three groups, according to how they model RNA evolution, and the theoretical differences in these classes are borne out by the different performance of the classes. The *All Pairs* models, which account for changes between all possible dinucleotides, are rarely chosen as the best-fit models. In part, this may be due to the composition of the test dataset of alignments, which is dominated by relatively short, well-conserved RNA genes. However, even in longer RNA genes these models do not describe RNA evolution well, which may be because the number of parameters required to model all of the non-canonical base pairs is disproportionate to their occurrence (~7%) in the RNA gene alignments.

The *Stable Pairs* models are chosen for approximately a third of the RNA gene alignments, and of the eight models in this group, 7G and 16C are predominant. These tend to be selected for the most conserved of the alignments, and the *Stable Pairs* models, and 7G in particular, are poor descriptions of evolution in RNA genes when they are not the best fitting model. This suggests that these models do not describe evolutionary events that are observable in RNA genes in general. It is not necessarily true that there should be general RNA evolutionary processes across all RNA genes, but I find that this does seem to be the case for *Stable Sets* models.

The models in the *Stable Sets* group, particularly the most general model in the group (16D), describe RNA evolution very well in most of the RNA gene alignments, even when not the absolute best-fit model. These models use the nucleotide, rather than dinucleotide, frequencies, and have parameters to describe the propensity for changes between Watson-Crick, wobble, and non-canonical base pairs. These models are perhaps the best at describing the multiple sources of constraint that potentially act on RNA genes. For example, the nucleotide frequencies might reflect the functional constraint on one arm of a microRNA precursor, with the propensity parameters describing the maintenance of the paired bases on the other arm.

Previous studies often reported that the most general models (either 7A or 16A) were the best descriptions of RNA evolution, and in particular that allowing double substitutions was important (Tillier and Collins, 1998; Higgs, 2000; Savill *et al.*, 2001),

in contrast to my results. This difference is likely due to differences in the datasets examined; previous analyses primarily used single rRNA alignments, and the mode of evolution in rRNA may well be atypical of RNA genes, because of its length and unique function. With a different dataset, with longer RNA genes and greater divergence, models that permit double substitutions may indeed be preferred, and this highlights the necessity of RNA model selection and the usefulness of the approach I demonstrate for comparing models across state space.

I do not examine the effect of using mixed models of loop and stem evolution, and it may be the case that the analysis of loops can mislead phylogenetic inference (Letsch *et al.*, 2010; Letsch and Kjer, 2011). These effects may underlie the results I describe in which the inferred tree differs from the established species tree. However, PHASE is currently unable to analyse only the stem regions of an RNA gene, and must also analyse the loop regions as part of a mixed model; given the potential problems with inference of loop regions, this ability would be a useful addition to PHASE. (This modification, although easy to state, is actually quite involved, because the current PHASE code relies quite heavily on using mixed model for RNA analysis.) It is also possible that although I use a variety of models to describe evolution in stems, none appropriately describe the evolution in some cases, and model misspecification can be problematic for phylogenetic inference. The model selection approach I describe allows for the evaluation of new models of evolution in stems, and adding these models to PHASE is possible but not simple, and this is another area of potential improvement for the software.

To perform phylogenetic analysis and model selection with RNA models it is necessary to partition an alignment, in order to appropriately apply nucleotide and dinucleotide models to the loops and stems, respectively, and thus a structure is required for use with an RNA model. So, if a structure is unavailable then an RNA model cannot be used for phylogenetic inference. This may not be a problem in practice, however, as structures can often be derived from homologous sequences with a known structure, or can be estimated by one of the many prediction methods. In a similar manner, to conduct model selection in a maximum likelihood framework a tree is required, which is problematic if one's goal is to infer a tree with an RNA model. However, a model can be chosen with a neighbour-joining tree, say, and then that model can then be used for a subsequent tree search.

The tree inference results that I report must be treated with caution, given the significant difference between many of the inferred trees and the species tree. Nonetheless,

there are clear differences between the trees inferred under different models, and it is interesting to note that the overlap between the DNA and *Stable Sets* models is relatively high. This suggests that the *Stable Sets* class, has more in common with the DNA models than the other RNA models, due to the incorporation of nucleotide, rather than dinucleotide, frequencies.

Chapter 6

Conclusions

In this thesis I set out to take a phylogenetic approach to improve our understanding of the evolution of RNA genes, and consequently the structure and function of ncRNA molecules. The first stage of any project is to evaluate whether a suitable dataset for analysis exists, or must be created. As has been concluded by others (Bateman *et al.*, 2011), there is no comprehensive resource for RNA data, and I found it necessary to generate alignments of RNA genes from existing genomic alignments. I chose this approach because it provides evolutionary, rather than structural, alignments (Kemena and Notredame, 2009), and allows for the analysis of genomic context, which is particularly important with RNA genes. RNA genes may be located near protein-coding genes, whose evolutionary signal must be distinguished from that of the RNA gene.

There are two publicly available sets of mammalian genomic alignments, the EPO and MultiZ datasets, and while the latter is widely used I found it much easier to work with the more recently developed EPO alignments. In programming terms, it is straightforward to extract up-to-date genome sequences from the EPO alignments with the Ensembl Perl API, and the longer block size negates the problem of how to join short blocks in MultiZ alignments. It was my original intention to base my analyses on the ‘high-coverage’ EPO-12 dataset, but it was a pleasant surprise to find that, after the same filtering steps that are required for the EPO-12 alignments, the larger EPO-35 dataset is of comparable quality. Clearly, the additional evolutionary information this represents is a boon for phylogenetic methods.

To generate and visualise alignments of RNA genes I developed the MARMOSET pipeline, as per the objective in section 1.5, and I believe that the creation of such alignments of RNA genes will be of use to others. If one wishes to perform phylogenetic analyses similar to those I conducted, the pipeline can take advantage of increasing

numbers of known RNA genes and genome sequences. The MARMOSET pipeline can also be used to generate alignments with a range of different properties, suitable for addressing a range of bioinformatic and biological questions. For example, if the alignments do not undergo any filtering steps, they can be used to generate datasets from different methods of genome alignment, such that the differences between alignment methods can be evaluated. Another interesting application might be the study of the gain and loss of RNA genes, to which genome alignments are particularly suited because flanking regions will be aligned. The strict filters I applied in Chapter 2 with respect to long insertions and gaps were designed to eliminate examples of gene gain and loss, as an unwanted complication for subsequent analyses, but the occasional example still crept in, as shown in Figure 2.4c.

I used the datasets of RNA gene alignments to evaluate *de novo* RNA gene prediction with RNAz and EvoFold because these programs had not previously been examined on a large positive set of known RNA genes; in addition, EvoFold had not been tested with an appropriately randomised negative dataset. It is interesting that the two programs, each reflecting a different methodology, detect similar true positives and dissimilar false positives, although the relatively poor performance of EvoFold, as an explicitly evolutionary approach, is somewhat disheartening. Other phylo-grammars have been shown to perform better than EvoFold (Bradley *et al.*, 2009b), but not to such a degree that I would expect a significant improvement for the RNA gene alignments in my study. EvoFold does use a relatively complex, 16-state, model of RNA evolution, and was trained on what is now a relatively old RNA dataset, such is the expansion of our knowledge about ncRNA; so an interesting future project might be to compare the performance of modified versions of EvoFold that use models from the different classes I describe in Chapter 5. My analysis of RNA gene prediction highlights an area where better knowledge of ncRNA evolution could be informative, and is thus complementary to the main objective of this work, the ability to evaluate DNA and RNA substitution models of evolution.

Until now, it has not been possible to conduct formal model selection with DNA and RNA models, so despite several papers indicating that RNA models better describe evolution in RNA genes than DNA models, it seems that researchers are unsure of the benefits when inferring species phylogenies with rRNA data, and still tend to use DNA models. There is a relatively large overlap between the trees inferred under DNA models and the best (*Stable Sets*) RNA models, so it is unlikely that these researchers have generated grossly incorrect trees (which I am sure would have been noticed anyway),

but certainly one would expect improvements in branch length or reliability by using models which better fit the data. I hope that the theory I developed to compare models across state space encourages the use of more appropriate models for phylogenetic inference of RNA genes. To aid this process, and corresponding to one of the objectives in section 1.5, the PHASE software has been modified to be more robust and includes a user-friendly Perl script to perform model selection with DNA and RNA models.

Two further objectives were to demonstrate, with a large number of alignments, that model selection with DNA and RNA models was useful (that is, to show that RNA models would sometimes fit the data better than DNA models); and, to examine the ability of different RNA models to describe evolution in RNA genes. The results showed an overwhelming preference for RNA models, in agreement with earlier work that studied rRNA alignments, and also suggested that certain types of model are more able to capture the patterns of ncRNA evolution than others. Although I used a relatively large set of alignments, its composition was skewed towards smaller genes such as miRNA and snoRNA, so the performance of the models may differ with a different set of data. This is to be expected, and is an important reason for performing rigorous model selection, enabled by my work on comparing models across state-space.

The ability to compare models with different numbers of states has applications beyond model selection for RNA data. For protein-coding genes, it enables the evaluation of whether data is best described at the nucleotide, amino acid, or codon level; this has been examined in the papers by Seo and Kishino (2008; 2009), but seems largely to have gone unnoticed by the research community, perhaps because of a lack of software for automating model selection, such as the popular jModelTest (Posada, 2008).

The proof in Chapter 4 also enables the evaluation of new types of evolutionary substitution model that recode alignments in a manner that is hypothesized to be biologically relevant, such as grouping amino acids based on physicochemical properties. Such models are not new, but being able to compare their performance, to each other and to existing models, might encourage the development of models that attempt to describe evolutionary patterns in a more nuanced and complex manner.

My intention with this work was to contribute to a better understanding of RNA gene evolution, and to create useful and reusable software for other researchers that share this aim. I have shown the usefulness of generating RNA gene alignments from genomic datasets, and I think the information in flanking sequences is particularly valuable for analysis of RNA genes. The ability to compare models of DNA and

RNA evolution in a standard framework, without the need for complicated simulations, removes a significant barrier to the phylogenetic inference of RNA genes with appropriate models. My results show that a single RNA model (16D) is the best (or a close second-best) fit for many RNA gene alignments, and the selection of this model is particularly interesting because it includes elements of both nucleotide and dinucleotide evolution, and thus may be capturing some of the complex sequence-level and structural-level constraints that exist in RNA genes. I believe that this represents a step towards more biologically-realistic, and useful, models of RNA evolution, that will enhance our knowledge of ncRNA and its variety of important biological roles.

References

- Adachi J and Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution* **42**(4):459–68.
- Adress KJ, Basilion JP, Klausner RD, Rouault TA, and Pardi A (1997) Structure and dynamics of the iron responsive element RNA: implications for binding of the RNA by iron regulatory binding proteins. *Journal of Molecular Biology* **274**(1):72–83.
- Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6):716–23.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**(17):3389–402.
- Anandam P, Torarinsson E, and Ruzzo WL (2009) Multiperm: shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies. *Bioinformatics* **25**(5):668–9.
- Anderson JW, Tataru P, Staines J, Hein J, and Lyngso R (2012) Evolving stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* **13**(1):78.
- Anisimova M and Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular Biology and Evolution* **26**(2):255–71.
- Arbiza L, Patricio M, Dopazo H, and Posada D (2011) Genome-wide heterogeneity of nucleotide substitution model fit. *Genome Biology and Evolution* **3**:896–908.
- Averof M, Rokas A, Wolfe KH, and Sharp PM (2000) Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* **287**(5456):1283–6.
- Avise JC (2006) *Evolutionary Pathways in Nature*. Cambridge University Press, Cambridge.
- Babak T, Blencowe BJ, and Hughes TR (2007) Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics* **8**:33.

- Bargelloni L, Ritchie PA, Patarnello T, Battaglia B, Lambert DM, and Meyer A (1994) Molecular evolution at subzero temperatures: mitochondrial and nuclear phylogenies of fishes from Antarctica (suborder Notothenioidei), and the evolution of antifreeze glycopeptides. *Molecular Biology and Evolution* **11**(6):854–63.
- Barquist L and Holmes I (2008) xREI: a phylo-grammar visualization webserver. *Nucleic Acids Research* **36**(Web Server issue):W65–9.
- Barry D and Hartigan JA (1987) Asynchronous distance between homologous DNA sequences. *Biometrics* **43**(2):261–76.
- Bateman A, Agrawal S, Birney E, Bruford EA, Bujnicki JM, Cochrane G, Cole JR, Dinger ME, Enright AJ, Gardner PP *et al.* (2011) RNAcentral: a vision for an international database of RNA sequences. *RNA* **17**(11):1941–6.
- Beitz E (2000) TEXshade: shading and labeling of multiple sequence alignments using LATEX2 epsilon. *Bioinformatics* **16**(2):135–9.
- Berezikov E (2011) Evolution of microRNA diversity and regulation in animals. *Nature Reviews. Genetics* **12**(12):846–60.
- Bernhart SH and Hofacker IL (2009) From consensus structure prediction to RNA gene finding. *Briefings in Functional Genomics & Proteomics* **8**(6):461–71.
- Bernhart SH, Hofacker IL, Will S, Gruber AR, and Stadler PF (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* **9**:474.
- Bida JP and Maher LJ (2012) Improved prediction of RNA tertiary structure with insights into native state dynamics. *RNA* **18**(3):385–93.
- Blackburne BP and Whelan S (2012) Measuring the distance between multiple sequence alignments. *Bioinformatics* **28**(4):495–502.
- Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* **14**(4):708–15.
- Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, and Taylor J (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology* **89**(19):10.1–21.
- Bofkin L and Goldman N (2007) Variation in evolutionary processes at different codon positions. *Molecular Biology and Evolution* **24**(2):513–21.
- Bonnet E, Wuyts J, Rouzé P, and Van de Peer Y (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20**(17):2911–7.

- Bouchard-Côté A, Sankararaman S, and Jordan MI (2012) Phylogenetic inference via sequential Monte Carlo. *Systematic Biology* **61**(4):579–93.
- Box GEP (1976) Science and statistics. *Journal of the American Statistical Association* **71**(356):791–9.
- Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, and Pachter L (2009a) Fast statistical alignment. *PLoS Computational Biology* **5**(5):e1000392.
- Bradley RK, Uzilov AV, Skinner ME, Bendaña YR, Barquist L, and Holmes I (2009b) Evolutionary modeling and prediction of non-coding RNAs in *Drosophila*. *PLoS ONE* **4**(8):e6478.
- Brosius J and Tiedge H (2004) RNomenclature. *RNA Biology* **1**(2):81–3.
- Brown CJ, Hendrich BD, Rupert JL, Lafrenière RG, Xing Y, Lawrence J, and Willard HF (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**(3):527–42.
- Bu D, Yu K, Sun S, Xie C, SkogerbøG, Miao R, Xiao H, Liao Q, Luo H, Zhao G *et al.* (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Research* **40**(D1):D210–D215.
- Burnham KP and Anderson DP (2004) Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods Research* **33**(2):261–304.
- Burnham KP and Anderson DR (2002) *Model Selection and Multimodel Inference*. Springer, New York, 2nd edition edition.
- Buschbom J and von Haeseler A (2005) Introduction to applications of the likelihood function in molecular evolution. In R Nielsen, ed., *Statistical Methods in Molecular Evolution*, pp. 25–44. Springer, New York, USA.
- Cannone JJ, Subramanian S, Schnare MN, Collett JR, D’Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3**:2.
- Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* **309**(5740):1559–63.
- Chang JM, Di Tommaso P, Taly JF, and Notredame C (2012) Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. *BMC Bioinformatics* **13 Suppl 4**:S1.

- Chen X and Tompa M (2010) Comparative assessment of methods for aligning multiple genome sequences. *Nature Biotechnology* **28**(6):567–72.
- Chomsky N (1959) On certain formal properties of grammars. *Information and Control* **2**(2):137–67.
- Coventry A, Kleitman DJ, and Berger B (2004) MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America* **101**(33):12102–7.
- Cuperus JT, Fahlgren N, and Carrington JC (2011) Evolution and functional diversification of miRNA genes. *The Plant Cell* **23**(2):431–42.
- Darty K, Denise A, and Ponty Y (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**(15):1974–5.
- Darwin C (1859) *On the Origin of Species by means of Natural Selection*. John Murray, London.
- Darzacq X, Jády BE, Verheggen C, Kiss AM, Bertrand E, and Kiss T (2002) Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs. *The EMBO Journal* **21**(11):2746–56.
- Dayhoff MO, Schwartz RM, and Orcutt BC (1978) A model of evolutionary change in proteins. In MO Dayhoff, ed., *Atlas of Protein Sequence and Structure*, volume 5, pp. 345–52. National Biomedical Research Foundation, Washington, D.C.
- Dayrat B (2003) The roots of phylogeny: how did Haeckel build his trees? *Systematic Biology* **52**(4):515–27.
- Delpont W, Scheffler K, and Seoighe C (2009) Models of coding sequence evolution. *Briefings in Bioinformatics* **10**(1):97–109.
- DeSalle R (2006) What's in a character? *Journal of Biomedical Informatics* **39**(1):6–17.
- Desper R and Gascuel O (2007) The minimum evolution distance-based approach to phylogenetic inference. In O Gascuel, ed., *Mathematics of Evolution and Phylogeny*, pp. 1–32. Oxford University Press, Oxford.
- Dimmic MW, Rest JS, Mindell DP, and Goldstein RA (2002) rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *Journal of Molecular Evolution* **55**(1):65–73.
- Do CB, Mahabhashyam MSP, Brudno M, and Batzoglou S (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Research* **15**(2):330–40.

- Dobzhansky T (1973) Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher* **35**:125–9.
- Dowell RD and Eddy SR (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* **5**:71.
- Dowell RD and Eddy SR (2006) Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics* **7**:400.
- Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Research* **40**(Database issue):D918–23.
- Durbin R, Eddy SR, Krogh A, and Mitchison G (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.
- Eddy SR (2001) Non-coding RNA genes and the modern RNA world. *Nature Reviews. Genetics* **2**(12):919–29.
- Edwards AWF (1972) *Likelihood*. Cambridge University Press, Cambridge.
- ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146):799–816.
- ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414):57–74.
- Evans D, Marquez SM, and Pace NR (2006) RNase P: interface of the RNA and protein worlds. *Trends in Biochemical Sciences* **31**(6):333–41.
- Faith DP (2006) Science and philosophy for molecular systematics: which is the cart and which is the horse? *Molecular Phylogenetics and Evolution* **38**(2):553–7.
- Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**(4):401–10.
- Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**(6):368–76.
- Felsenstein J (2004) *Inferring Phylogenies*. Sinauer, Sunderland, Massachusetts.
- Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. *PHYLIP Package. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.* .
- Felsenstein J and Churchill GA (1996) A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution* **13**(1):93–104.

- Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology* **20**(4):406–16.
- Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S *et al.* (2012) Ensembl 2012. *Nucleic Acids Research* **40**(Database issue):D84–90.
- Freyhult EK, Bollback JP, and Gardner PP (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Research* **17**(1):117–25.
- Friedman RC, Farh KKH, Burge CB, and Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* **19**(1):92–105.
- Gardner PP (2009) The use of covariance models to annotate RNAs in whole genomes. *Briefings in Functional Genomics & Proteomics* **8**(6):444–50.
- Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR *et al.* (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Research* **39**(Database issue):D141–5.
- Gardner PP and Giegerich R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* **5**:140.
- Gardner PP, Wilm A, and Washietl S (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research* **33**(8):2433–9.
- Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* **14**(7):685–95.
- Gelfman S, Burstein D, Penn O, Savchenko A, Amit M, Schwartz S, Pupko T, and Ast G (2012) Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome Research* **22**(1):35–50.
- Gesell T and Washietl S (2008) Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics* **9**:248.
- Gibson A, Gowri-Shankar V, Higgs PG, and Rattray M (2005) A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. *Molecular Biology and Evolution* **22**(2):251–64.
- Gilbert W (1986) The RNA World. *Nature* **319**:618.
- Girard A, Sachidanandam R, Hannon GJ, and Carmell MA (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**(7099):199–202.
- Goldman N (1993) Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**(2):182–98.

- Goldman N, Thorne JL, and Jones DT (1996) Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *Journal of Molecular Biology* **263**(2):196–208.
- Goldman N and Whelan S (2000) Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Molecular Biology and Evolution* **17**(6):975–8.
- Goldman N and Whelan S (2002) A novel use of equilibrium frequencies in models of sequence evolution. *Molecular Biology and Evolution* **19**(11):1821–31.
- Goldman N and Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**(5):725–36.
- Gorodkin J and Hofacker IL (2011) From structure prediction to genomic screens for novel non-coding RNAs. *PLoS Computational Biology* **7**(8):e1002100.
- Gorodkin J, Hofacker IL, Torarinsson E, Yao Z, Havgaard JH, and Ruzzo WL (2010) De novo prediction of structured RNAs from genomic sequences. *Trends in Biotechnology* **28**(1):9–19.
- Gowri-Shankar V and Rattray M (2006) On the correlation between composition and site-specific evolutionary rate: implications for phylogenetic inference. *Molecular Biology and Evolution* **23**(2):352–64.
- Gowri-Shankar V and Rattray M (2007) A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model. *Molecular Biology and Evolution* **24**(6):1286–99.
- Granneman S and Baserga SJ (2004) Ribosome biogenesis: of knobs and RNA processing. *Experimental Cell Research* **296**(1):43–50.
- Griffiths-Jones S (2007) Annotating noncoding RNA genes. *Annual Review of Genomics and Human Genetics* **8**:279–98.
- Gruber AR, Bernhart SH, Hofacker IL, and Washietl S (2008) Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics* **9**:122.
- Gruber AR, Findeiß S, Washietl S, Hofacker IL, and Stadler PF (2010) RNAz 2.0: improved noncoding RNA detection. *Pacific Symposium on Biocomputing* **15**:69–79.
- Guerrier-Takada C, Gardiner K, Marsh T, Pace N, and Altman S (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**(3 Pt 2):849–57.
- Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, and Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**(3):307–21.

- Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**(7235):223–7.
- Harris RS (2007) *Improved pairwise alignment of genomic DNA*. The Pennsylvania State University.
- Hasegawa M, Kishino H, and Yano Ta (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**(2):160–74.
- Havgaard JH, Torarinsson E, and Gorodkin J (2007) Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Computational Biology* **3**(10):1896–908.
- He L and Hannon GJ (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews. Genetics* **5**(7):522–31.
- Heimberg AM, Cowper-Sal Lari R, Sémon M, Donoghue PCJ, and Peterson KJ (2010) microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proceedings of the National Academy of Sciences of the United States of America* **107**(45):19379–83.
- Helfenbein KG and DeSalle R (2005) Falsifications and corroborations: Karl Popper’s influence on systematics. *Molecular Phylogenetics and Evolution* **35**(1):271–80.
- Herbig A and Nieselt K (2011) nocoRNAC: characterization of non-coding RNAs in prokaryotes. *BMC Bioinformatics* **12**(1):40.
- Higgs PG (2000) RNA secondary structure: physical and computational aspects. *Quarterly Reviews of Biophysics* **33**(3):199–253.
- Hiller M, Schaar BT, and Bejerano G (2012) Hundreds of conserved non-coding genomic regions are independently lost in mammals. *Nucleic Acids Research* p. Epub.
- Hofacker IL (2003) Vienna RNA secondary structure server. *Nucleic Acids Research* **31**(13):3429–31.
- Hofacker IL, Fekete M, and Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. *Journal of Molecular Biology* **319**(5):1059–66.
- Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, and Schuster P (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly* **125**(2):167–88.
- Holbrook SR (2008) Structural principles from large RNAs. *Annual Review of Biophysics* **37**:445–64.

- Hudelot C, Gowri-Shankar V, Jow H, Rattray M, and Higgs PG (2003) RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Molecular Phylogenetics and Evolution* **28**(2):241–52.
- Huelsenbeck JP and Rannala B (1997) Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**(5310):227–32.
- Huelsenbeck JP and Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**(8):754–5.
- Huelsenbeck JP and Ronquist F (2005) Bayesian analysis of molecular evolution using MrBayes. In R Nielsen, ed., *Statistical Methods in Molecular Evolution*, pp. 183–232. Springer, New York, USA.
- Huelsenbeck JP, Ronquist F, Nielsen R, and Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**(5550):2310–4.
- Hulsen T, de Vlieg J, and Alkema W (2008) BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics* **9**:488.
- Jayaswal V, Ababneh F, Jermiin LS, and Robinson J (2011) Reducing model complexity of the general markov model of evolution. *Molecular Biology and Evolution* **28**(11):3045–59.
- Jayaswal V, Jermiin LS, and Robinson J (2005) Estimation of phylogeny using a general markov model. *Evolutionary Bioinformatics Online* **1**:62–80.
- Jeggari A, Marks DS, and Larsson E (2012) miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics* **28**(15):2062–3.
- Jones DT, Taylor WR, and Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences: CABIOS* **8**(3):275–82.
- Jones DT, Taylor WR, and Thornton JM (1994) A mutation data matrix for transmembrane proteins. *FEBS Letters* **339**(3):269–75.
- Jow H, Hudelot C, Rattray M, and Higgs PG (2002) Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. *Molecular Biology and Evolution* **19**(9):1591–601.
- Jukes TH and Cantor CR (1969) Evolution of protein molecules. In HN Munro, ed., *Mammalian protein metabolism*, volume 3, pp. 21–132. Academic Press, New York.
- Katoh K and Toh H (2008) Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics* **9**:212.

- Keane TM, Creevey CJ, Pentony MM, Naughton TJ, and McInerney JO (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evolutionary Biology* **6**:29.
- Kedersha NL and Rome LH (1986) Isolation and characterization of a novel ribonucleoprotein particle: large structures contain a single species of small RNA. *The Journal of Cell Biology* **103**(3):699–709.
- Kelchner SA and Thomas MA (2007) Model use in phylogenetics: nine key questions. *Trends in Ecology & Evolution* **22**(2):87–94.
- Kemena C and Notredame C (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* **25**(19):2455–65.
- Kendall MG and Stuart A (1973) *The Advanced Theory of Statistics*, volume 2. Griffin, London, 3rd edition edition.
- Kent WJ (2002) BLAT—the BLAST-like alignment tool. *Genome Research* **12**(4):656–64.
- Kent WJ, Baertsch R, Hinrichs A, Miller W, and Haussler D (2003) Evolution’s cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America* **100**(20):11484–9.
- Kim J (1996) General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Systematic Biology* **4**(3):363–74.
- Kim J and Sanderson MJ (2008) Penalized likelihood phylogenetic inference: bridging the parsimony-likelihood gap. *Systematic Biology* **57**(5):665–74.
- Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, and Davidson GS (2001) A gene expression map for *Caenorhabditis elegans*. *Science* **293**(5537):2087–92.
- Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**(2):111–20.
- Kishore S and Stamm S (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* **311**(5758):230–2.
- Kiss T (2001) Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *The EMBO Journal* **20**(14):3617–22.
- Klein RJ and Eddy SR (2003) RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics* **4**:44.

- Klosterman PS, Uzilov AV, Bendaña YR, Bradley RK, Chao S, Kosiol C, Goldman N, and Holmes I (2006) XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics* **7**:428.
- Knudsen B and Hein J (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research* **31**(13):3423–8.
- Kolaczkowski B and Thornton JW (2007) Effects of branch length uncertainty on Bayesian posterior probabilities for phylogenetic hypotheses. *Molecular Biology and Evolution* **24**(9):2108–18.
- Kosiol C, Bofkin L, and Whelan S (2006) Phylogenetics by likelihood: evolutionary modeling as a tool for understanding the genome. *Journal of Biomedical Informatics* **39**(1):51–61.
- Kozomara A and Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research* **39**(Database issue):D152–7.
- Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, and Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**(9):3100–8.
- Lanave C, Preparata G, Saccone C, and Serio G (1984) A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* **20**(1):86–93.
- Lanfear R, Calcott B, Ho SYW, and Guindon S (2012) Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* **29**(6):1695–701.
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**(21):2947–8.
- Le SQ and Gascuel O (2008) An improved general amino acid replacement matrix. *Molecular Biology and Evolution* **25**(7):1307–20.
- Letsch HO and Kjer KM (2011) Potential pitfalls of modelling ribosomal RNA data in phylogenetic tree reconstruction: evidence from case studies in the Metazoa. *BMC Evolutionary Biology* **11**:146.
- Letsch HO, Kück P, Stocsits RR, and Misof B (2010) The impact of rRNA secondary structure consideration in alignment and tree reconstruction: simulated data and a case study on the phylogeny of hexapods. *Molecular Biology and Evolution* **27**(11):2507–21.
- Liò P and Goldman N (1998) Models of molecular evolution and phylogeny. *Genome Research* **8**(12):1233–44.

- Lopez P, Casane D, and Philippe H (2002) Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution* **19**(1):1–7.
- Lowe TM and Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**(5):955–64.
- Löytynoja A and Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America* **102**(30):10557–62.
- Löytynoja A and Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**(5883):1632–5.
- Machado-Lima A, del Portillo HA, and Durham AM (2008) Computational methods in noncoding RNA research. *Journal of Mathematical Biology* **56**(1-2):15–49.
- Maddison WP and Maddison DR (2011) Mesquite: a modular system for evolutionary analysis. Version 2.75.
- Mallatt J, Craig CW, and Yoder MJ (2010) Nearly complete rRNA genes assembled from across the metazoan animals: effects of more taxa, a structure-based alignment, and paired-sites evolutionary models on phylogeny reconstruction. *Molecular Phylogenetics and Evolution* **55**(1):1–17.
- Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, Keefe D, Schwartz AS, Hou M *et al.* (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Research* **17**(6):760–74.
- Markham NR and Zuker M (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods in Molecular Biology* **453**:3–31.
- Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, and Turner DH (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America* **101**(19):7287–92.
- Mattick JS (2009) The genetic signatures of noncoding RNAs. *PLoS Genetics* **5**(4):e1000459.
- Mattick JS and Makunin IV (2006) Non-coding RNA. *Human Molecular Genetics* **15 Spec No 1**:R17–29.
- Menzel P, Gorodkin J, and Stadler PF (2009) The tedious task of finding homologous noncoding RNA genes. *RNA* **15**(12):2075–82.
- Mercer TR, Dinger ME, and Mattick JS (2009) Long non-coding RNAs: insights into functions. *Nature Reviews. Genetics* **10**(3):155–9.

- Meyer IM (2007) A practical guide to the art of RNA gene prediction. *Briefings in Bioinformatics* **8**(6):396–414.
- Minin V, Abdo Z, Joyce P, and Sullivan J (2003) Performance-based selection of likelihood models for phylogeny estimation. *Systematic Biology* **52**(5):674–83.
- Mituyama T, Yamada K, Hattori E, Okida H, Ono Y, Terai G, Yoshizawa A, Komori T, and Asai K (2009) The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Research* **37**(Database issue):D89–92.
- Morozova N, Zinovyev A, Nonne N, Pritchard LL, Gorban AN, and Harel-Bellan A (2012) Kinetic signatures of microRNA modes of action. *RNA* **18**(9):1635–55.
- Morrison D (2009a) A framework for phylogenetic sequence alignment. *Plant Systematics and Evolution* **282**(3):127–49.
- Morrison DA (2009b) Why would phylogeneticists ignore computerized sequence alignment? *Systematic Biology* **58**(1):150–8.
- Muse SV (1995) Evolutionary analyses of DNA sequences subject to constraints of secondary structure. *Genetics* **139**(3):1429–39.
- Muse SV and Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**(5):715–24.
- Nawrocki EP, Kolbe DL, and Eddy SR (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**(10):1335–7.
- Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**(3):443–53.
- Newcombe RG (1998) Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17**(8):857–72.
- Noller HF (2005) RNA structure: reading the ribosome. *Science* **309**(5740):1508–14.
- Novák A, Miklós I, LyngsøR, and Hein J (2008) StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics* **24**(20):2403–4.
- Oulas A, Reczko M, and Poirazi P (2009) MicroRNAs and cancer—the search begins! *IEEE Transactions on Information Technology in Biomedicine* **13**(1):67–77.
- Outlaw DC, Voelker G, Mila B, and Girman DJ (2003) Evolution of long-distance migration in and historical biogeography of Catharus thrushes: a molecular phylogenetic approach. *The Auk* **120**(2):299–310.

- Pagel M and Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology* **53**(4):571–81.
- Pagel M and Meade A (2008) Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **363**(1512):3955–64.
- Pang KC, Frith MC, and Mattick JS (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in Genetics* **22**(1):1–5.
- Pang KC, Stephen S, Dinger ME, Engström PG, Lenhard B, and Mattick JS (2007) RNAdb 2.0—an expanded database of mammalian non-coding RNAs. *Nucleic Acids Research* **35**(Database issue):D178–82.
- Parker BJ, Moltke I, Roth A, Washietl S, Wen J, Kellis M, Breaker R, and Pedersen JS (2011) New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Research* **21**(11):1929–43.
- Paten B, Diekhans M, Earl D, John JS, Ma J, Suh B, and Haussler D (2011a) Cactus graphs for genome comparisons. *Journal of Computational Biology* **18**(3):469–81.
- Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, and Haussler D (2011b) Cactus: algorithms for genome multiple sequence alignment. *Genome Research* **21**(9):1512–28.
- Paten B, Herrero J, Beal K, and Birney E (2009) Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics* **25**(3):295–301.
- Paten B, Herrero J, Beal K, Fitzgerald S, and Birney E (2008a) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research* **18**(11):1814–28.
- Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, and Birney E (2008b) Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Research* **18**(11):1829–43.
- Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, and Haussler D (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Computational Biology* **2**(4):e33.
- Persson H, Kvist A, Vallon-Christersson J, Medstrand P, Borg A, and Rovira C (2009) The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. *Nature Cell Biology* **11**(10):1268–71.

- Pol D (2004) Empirical problems of the hierarchical likelihood ratio test for model selection. *Systematic Biology* **53**(6):949–62.
- Poole AM, Jeffares DC, and Penny D (1998) The path from the RNA world. *Journal of Molecular Evolution* **46**(1):1–17.
- Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* **25**(7):1253–6.
- Posada D and Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology* **53**(5):793–808.
- Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, and Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**(21):7188–96.
- Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J *et al.* (2012) The Pfam protein families database. *Nucleic Acids Research* **40**(Database issue):D290–301.
- Raasch P, Schmitz U, Patenge N, Vera J, Kreikemeyer B, and Wolkenhauer O (2010) Non-coding RNA detection methods combined to improve usability, reproducibility and precision. *BMC Bioinformatics* **11**(1):491.
- Rannala B and Yang Z (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution* **43**(3):304–11.
- Raxworthy CJ, Forstner MRJ, and Nussbaum RA (2002) Chameleon radiation by oceanic dispersal. *Nature* **415**(6873):784–7.
- Reeder J, Höchsmann M, Rehmsmeier M, VoßB, and Giegerich R (2006) Beyond Mfold: recent advances in RNA bioinformatics. *Journal of Biotechnology* **124**(1):41–55.
- Ren F, Tanaka H, and Yang Z (2005) An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Systematic Biology* **54**(5):808–18.
- Rice P, Longden I, and Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics* **16**(6):276–7.
- Rinn JL and Chang HY (2012) Genome regulation by long noncoding RNAs. *Annual Review of Biochemistry* **81**:145–66.
- Rivas E and Eddy SR (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **16**(7):583–605.

- Rivas E and Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2**:8.
- Rivas E, Lang R, and Eddy SR (2012) A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA* **18**(2):193–212.
- Robinson DF and Foulds LR (1981) Comparison of phylogenetic trees. *Mathematical Biosciences* **53**(1-2):131–47.
- Robinson DM, Jones DT, Kishino H, Goldman N, and Thorne JL (2003) Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution* **20**(10):1692–704.
- Rodrigue N, Lartillot N, Bryant D, and Philippe H (2005) Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* **347**(2):207–17.
- Rodrigue N, Lartillot N, and Philippe H (2008) Bayesian comparisons of codon substitution models. *Genetics* **180**(3):1579–91.
- Rodrigue N, Philippe H, and Lartillot N (2006) Assessing site-interdependent phylogenetic models of sequence evolution. *Molecular Biology and Evolution* **23**(9):1762–75.
- Rodríguez F, Oliver JL, Marín A, and Medina JR (1990) The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology* **142**(4):485–501.
- Romiguier J, Ranwez V, Douzery EJP, and Galtier N (2010) Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Research* **20**(8):1001–9.
- Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, and Huelsenbeck JP (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**(3):539–42.
- Rose D, Hackermüller J, Washietl S, Reiche K, Hertel J, Findeiß S, Stadler PF, and Prohaska SJ (2007) Computational RNomics of drosophilids. *BMC Genomics* **8**:406.
- Rother M, Rother K, Puton T, and Bujnicki JM (2011) RNA tertiary structure prediction with ModeRNA. *Briefings in Bioinformatics* **12**(6):601–13.
- Rzhetsky A (1995) Estimating substitution rates in ribosomal RNA genes. *Genetics* **141**(2):771–83.
- Sahraeian SME and Yoon BJ (2011) PicXAA-R: efficient structural alignment of multiple RNA sequences using a greedy approach. *BMC Bioinformatics* **12 Suppl 1**:S38.

- Saito Y, Sato K, and Sakakibara Y (2010) Robust and accurate prediction of noncoding RNAs from aligned sequences. *BMC Bioinformatics* **11 Suppl 7**:S3.
- Saitou N and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**(4):406–25.
- Sakakibara Y, Brown M, Hughey R, Mian IS, Sjölander K, Underwood RC, and Haussler D (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research* **22**(23):5112–20.
- Savill NJ, Hoyle DC, and Higgs PG (2001) RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics* **157**(1):399–411.
- Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lapalainen T, Mailund T, Marques-Bonet T *et al.* (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**(7388):169–75.
- Schöniger M and von Haeseler A (1994) A stochastic model for the evolution of auto-correlated DNA sequences. *Molecular Phylogenetics and Evolution* **3**(3):240–7.
- Schöniger M and von Haeseler A (1999) Toward assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models. *Journal of Molecular Evolution* **49**(5):691–8.
- Schwach F, Moxon S, Moulton V, and Dalmay T (2009) Deciphering the diversity of small RNAs in plants: the long and short of it. *Briefings in Functional Genomics & Proteomics* **8**(6):472–81.
- Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, and Miller W (2003) Human-mouse alignments with BLASTZ. *Genome Research* **13**(1):103–7.
- Sellers PH (1974) On the theory and computation of evolutionary distance. *SIAM Journal on Applied Mathematics* **26**(4):787–93.
- Seo TK and Kishino H (2008) Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. *Systematic Biology* **57**(3):367–77.
- Seo TK and Kishino H (2009) Statistical comparison of nucleotide, amino acid, and codon substitution models for evolutionary analysis of protein-coding sequences. *Systematic Biology* **58**(2):199–210.
- Shen Y, Lv Y, Huang L, Liu W, Wen M, Tang T, Zhang R, Hungate E, Shi S, and Wu CI (2011) Testing hypotheses on the rate of molecular evolution in relation to gene expression using microRNAs. *Proceedings of the National Academy of Sciences of the United States of America* **108**(38):15942–7.

- Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Systematic Biology* **51**(3):492–508.
- Shimodaira H and Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* **16**(8):1114–6.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**(8):1034–50.
- Siepel A and Haussler D (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology* **11**(2-3):413–28.
- Smith NGC, Webster MT, and Ellegren H (2003) A low rate of simultaneous double-nucleotide mutations in primates. *Molecular Biology and Evolution* **20**(1):47–53.
- Sokal RR and Sneath PHA (1963) *Principles of Numerical Taxonomy*. W.H. Freeman, San Francisco.
- Song D, Yang Y, Yu B, Zheng B, Deng Z, Lu BL, Chen X, and Jiang T (2009) Computational prediction of novel non-coding RNAs in *Arabidopsis thaliana*. *BMC Bioinformatics* **10 Suppl 1**:S36.
- Sperling EA, Vinther J, Moy VN, Wheeler BM, Sémon M, Briggs DEG, and Peterson KJ (2009) MicroRNAs resolve an apparent conflict between annelid systematics and their fossil record. *Proceedings of the Royal Society B: Biological Sciences* **276**(1677):4315–22.
- Squartini F and Arndt PF (2008) Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Molecular Biology and Evolution* **25**(12):2525–35.
- Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**(21):2688–90.
- Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN *et al.* (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**(7167):219–32.
- Steel M (2005) Should phylogenetic models be trying to "fit an elephant"? *Trends in Genetics* **21**(6):307–9.
- Steel M and Penny D (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution* **17**(6):839–50.
- Steigle S, Huber W, Stocsits C, Stadler PF, and Nieselt K (2007) Comparative analysis of structured RNAs in *S. cerevisiae* indicates a multitude of different functions. *BMC Biology* **5**:25.

- Suchard MA and Redelings BD (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* **22**(16):2047–8.
- Sullivan J and Joyce P (2005) Model selection in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* **36**:445–66.
- Suutari M, Majaneva M, Fewer DP, Voirin B, Aiello A, Friedl T, Chiarello AG, and Blomster J (2010) Molecular evidence for a diverse green algal community growing in the hair of sloths and a specific association with *Trichophilus welckeri* (Chlorophyta, Ulvophyceae). *BMC Evolutionary Biology* **10**:86.
- Taft RJ, Pheasant M, and Mattick JS (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays* **29**(3):288–99.
- Tamura K and Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10**(3):512–26.
- Taufer M, Licon A, Araiza R, Mireles D, van Batenburg FHD, Gulyaev AP, and Leung MY (2009) PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. *Nucleic Acids Research* **37**(Database issue):D127–35.
- Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**:57–86.
- Taylor WR (1997) Residual colours: a proposal for aminochromography. *Protein Engineering* **10**(7):743–6.
- Telford MJ, Wise MJ, and Gowri-Shankar V (2005) Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: examples from the bilateria. *Molecular Biology and Evolution* **22**(4):1129–36.
- Thompson JD, Koehl P, Ripp R, and Poch O (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins* **61**(1):127–36.
- Thorne JL, Kishino H, and Felsenstein J (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution* **33**(2):114–24.
- Thorne JL, Kishino H, and Felsenstein J (1992) Inching toward reality: an improved likelihood model of sequence evolution. *Journal of Molecular Evolution* **34**(1):3–16.
- Tillier ERM and Collins RA (1995) Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. *Molecular Biology and Evolution* **12**(1):7–15.

- Tillier ERM and Collins RA (1998) High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics* **148**(4):1993–2002.
- Torarinsson E, Yao Z, Wiklund ED, Bramsen JB, Hansen C, Kjems J, Tommerup N, Ruzzo WL, and Gorodkin J (2008) Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Research* **18**(2):242–51.
- Trajkovski M, Hausser J, Soutschek J, Bhat B, Akin A, Zavolan M, Heim MH, and Stoffel M (2011) MicroRNAs 103 and 107 regulate insulin sensitivity. *Nature* **474**(7353):649–53.
- Tuffley C and Steel M (1997) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology* **59**(3):581–607.
- Uzilov AV, Keegan JM, and Mathews DH (2006) Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics* **7**:173.
- Uzzell T and Corbin KW (1971) Fitting discrete probability distributions to evolutionary events. *Science* **172**(988):1089–96.
- Valadkhan S (2005) snRNAs as the catalysts of pre-mRNA splicing. *Current Opinion in Chemical Biology* **9**(6):603–8.
- Valastyan S, Reinhardt F, Benaich N, Calogrias D, Szász AM, Wang ZC, Brock JE, Richardson AL, and Weinberg RA (2009) A pleiotropically acting microRNA, miR-31, inhibits breast cancer metastasis. *Cell* **137**(6):1032–46.
- van Batenburg FHD, Gulyaev AP, Pleij CWA, Ng J, and Oliehoek J (2000) Pseudo-Base: a database with RNA pseudoknots. *Nucleic Acids Research* **28**(1):201–4.
- Varadarajan A, Bradley RK, and Holmes IH (2008) Tools for simulating evolution of aligned genomic regions with integrated parameter estimation. *Genome Biology* **9**(10):R147.
- von Reumont BM, Meusemann K, Szucsich NU, Dell’Ampio E, Gowri-Shankar V, Bartel D, Simon S, Letsch HO, Stocsits RR, Luan Yx *et al.* (2009) Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships. *BMC Evolutionary Biology* **9**:119.
- Walczak R, Westhof E, Carbon P, and Krol A (1996) A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *RNA* **2**(4):367–79.

- Wang AX, Ruzzo WL, and Tompa M (2007a) How accurately is ncRNA aligned within whole-genome multiple alignments? *BMC Bioinformatics* **8**:417.
- Wang HC, Spencer M, Susko E, and Roger AJ (2007b) Testing for covarion-like evolution in protein sequences. *Molecular Biology and Evolution* **24**(1):294–305.
- Wang KC and Chang HY (2011) Molecular mechanisms of long noncoding RNAs. *Molecular Cell* **43**(6):904–14.
- Wang PPS and Ruvinsky I (2012) Family size and turnover rates among several classes of small non-protein-coding RNA genes in *Caenorhabditis* nematodes. *Genome Biology and Evolution* **4**(4):565–74.
- Washietl S, Findeiß S, Müller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, and Goldman N (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* **17**:578–94.
- Washietl S and Hofacker IL (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *Journal of Molecular Biology* **342**(1):19–30.
- Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, and Stadler PF (2005a) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nature Biotechnology* **23**(11):1383–90.
- Washietl S, Hofacker IL, and Stadler PF (2005b) Fast and reliable prediction of non-coding RNAs. *Proceedings of the National Academy of Sciences of the United States of America* **102**(7):2454–9.
- Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A *et al.* (2007) Structured RNAs in the ENCODE selected regions of the human genome. *Genome Research* **17**(6):852–64.
- Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, Sasaki H, Minami N, and Imai H (2006) Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes & Development* **20**(13):1732–43.
- Waterhouse AM, Procter JB, Martin DMA, Clamp M, and Barton GJ (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**(9):1189–91.
- Waters LS and Storz G (2009) Regulatory RNAs in bacteria. *Cell* **136**(4):615–28.
- Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N *et al.* (2007) Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Research* **35**(14):4809–19.

- Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, and Breaker RR (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biology* **11**(3):R31.
- Whelan S (2007) New approaches to phylogenetic tree search and their application to large numbers of protein alignments. *Systematic Biology* **56**(5):727–40.
- Whelan S (2008a) The genetic code can cause systematic bias in simple phylogenetic models. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **363**(1512):4003–11.
- Whelan S (2008b) Spatial and temporal heterogeneity in nucleotide sequence evolution. *Molecular Biology and Evolution* **25**(8):1683–94.
- Whelan S and Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* **18**(5):691–9.
- Whelan S and Goldman N (2004) Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics* **167**(4):2027–43.
- Whelan S, Liò P, and Goldman N (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics* **17**(5):262–72.
- Will S, Joshi T, Hofacker IL, Stadler PF, and Backofen R (2012) LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA* **18**(5):900–14.
- Will S, Reiche K, Hofacker IL, Stadler PF, and Backofen R (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology* **3**(4):e65.
- Wilson EB (1927) Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**(158):209–12.
- Woese CR and Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America* **74**(11):5088–90.
- Wong KM, Suchard MA, and Huelsenbeck JP (2008) Alignment uncertainty and genomic analysis. *Science* **319**(5862):473–6.
- Workman C and Krogh A (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Research* **27**(24):4816–22.
- Yang Z (1994a) Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* **39**(1):105–11.

- Yang Z (1994b) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39**(3):306–14.
- Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution* **11**(9):367–72.
- Yang Z (2006) *Computational Molecular Evolution*. Oxford University Press.
- Yang Z, Goldman N, and Friday A (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution* **11**(2):316–24.
- Yang Z, Nielsen R, and Hasegawa M (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution* **15**(12):1600–11.
- Yang Z and Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Molecular Biology and Evolution* **14**(7):717–24.
- Yang Z and Rannala B (2012) Molecular phylogenetics: principles and practice. *Nature Reviews. Genetics* **13**(5):303–14.
- Yao Z, Barrick J, Weinberg Z, Neph S, Breaker R, Tompa M, and Ruzzo WL (2007) A computational pipeline for high-throughput discovery of cis-regulatory noncoding RNA in prokaryotes. *PLoS Computational Biology* **3**(7):e126.
- Yao Z, Weinberg Z, and Ruzzo WL (2006) CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics* **22**(4):445–52.
- Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, and Noller HF (2001) Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292**(5518):883–96.
- Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution* **39**(3):315–29.
- Zhong C, Andrews J, and Zhang S (2012) Discovering non-coding RNA elements in drosophila 3' untranslated regions. *2nd IEEE International Conference on Computational Advances in Bio and Medical Sciences* .
- Zuker M and Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research* **9**(1):133–48.

Appendix A

Mapping eigenvalues from compound to distinct models

Equation 4.11 in Chapter 4 expresses the eigenvalues of distinct-state matrices in terms of parameters from compound-state matrices and their eigenvalues. The derivation of these equations can be demonstrated for a small example, and extension to larger matrices can be inferred because eigendecomposition of larger matrices is an iterative process, and because the matrix \mathbf{Q}_D is inherently symmetrical by definition.

Let there be two compound states c_1 and c_2 , with associated distinct states $d_{1,1}, d_{1,2}$, and $d_{2,1}$. The rate matrices for the compound and distinct models are (from equations 4.1 and 4.2), respectively:

$$\mathbf{Q}_C = \begin{pmatrix} -s\pi_2 & s\pi_2 \\ s\pi_1 & -s\pi_1 \end{pmatrix}$$

and

$$\mathbf{Q}_D = \begin{pmatrix} -\rho\pi_{d_{1,2}} - s\pi_{c_2} & \rho\pi_{d_{1,2}} & s\pi_{c_2} \\ \rho\pi_{d_{1,1}} & -\rho\pi_{d_{1,1}} - s\pi_{c_2} & s\pi_{c_2} \\ s\pi_{d_{1,1}} & s\pi_{d_{1,2}} & -s\pi_{c_1} \end{pmatrix}$$

where $s = s_{c_1c_2}$ and $\rho = \rho_1$ in order to eliminate redundant notation. The eigenvalues

of \mathbf{Q}_C are:

$$\lambda_1 = -s, \lambda_2 = 0$$

The characteristic polynomial of \mathbf{Q}_D , derived by multiplying the matrix by $-\mu\mathbf{I}$ (where \mathbf{I} is the identity matrix) and calculating the determinant gives a cubic equation in μ :

$$\begin{aligned} &(-\rho\pi_{d_{1,2}} - s\pi_{c_2} - \mu)(-\rho\pi_{d_{1,1}} - s\pi_{c_2} - \mu)(-s\pi_{c_1} - \mu) + \\ &\quad \rho s^2 \pi_{d_{1,1}} \pi_{d_{1,2}} \pi_{c_2} + \\ &\quad \rho s^2 \pi_{d_{1,1}} \pi_{d_{1,2}} \pi_{c_2} + \\ &\quad \rho s^2 \pi_{d_{1,1}} \pi_{d_{1,1}} \pi_{c_2} + s^3 \pi_{d_{1,1}} \pi_{c_2} \pi_{c_2} + \mu s^2 \pi_{d_{1,1}} \pi_{c_2} + \\ &\quad \rho^2 s \pi_{d_{1,1}} \pi_{d_{1,2}} \pi_{c_1} + \mu \rho^2 \pi_{d_{1,1}} \pi_{d_{1,2}} + \\ &\quad \rho s^2 \pi_{d_{1,2}} \pi_{d_{1,2}} \pi_{c_2} + s^3 \pi_{d_{1,2}} \pi_{c_2} \pi_{c_2} + \mu s^2 \pi_{d_{1,2}} \pi_{c_2} = \end{aligned}$$

$$\begin{aligned} &(\rho^2 \pi_{d_{1,1}} \pi_{d_{1,2}} + \rho s \pi_{d_{1,1}} \pi_{c_2} + \mu \rho \pi_{d_{1,1}})(-s\pi_{c_1} - \mu) + \\ &\quad (\rho s \pi_{d_{1,2}} \pi_{c_2} + s^2 \pi_{c_2} \pi_{c_2} + \mu s \pi_{c_2})(-s\pi_{c_1} - \mu) + \\ &\quad (\mu \rho \pi_{d_{1,2}} + \mu s \pi_{c_2} + \mu^2)(-s\pi_{c_1} - \mu) + \\ &\quad \mu(\rho^2 \pi_{d_{1,1}} \pi_{d_{1,2}} + s^2 \pi_{c_1} \pi_{c_2}) + \\ &\quad \rho^2 s \pi_{d_{1,1}} \pi_{d_{1,2}} \pi_{c_1} + \rho s^2 \pi_{c_1} \pi_{c_1} \pi_{c_2} + s^3 \pi_{c_1} \pi_{c_2} \pi_{c_2} = \end{aligned}$$

$$-\mu^3 +$$

$$-\mu^2(\rho\pi_{c_1} + s\pi_{c_2} + s) +$$

$$-\mu(\rho^2 \pi_{d_{1,1}} \pi_{d_{1,2}} + \rho s \pi_{c_1} + s^2 \pi_{c_2} + s^2 \pi_{c_1} \pi_{c_2}) +$$

$$-(\rho^2 s \pi_{d_{1,1}} \pi_{d_{1,2}} \pi_{c_1} + \rho s^2 \pi_{c_1} \pi_{c_1} \pi_{c_2} + s^3 \pi_{c_1} \pi_{c_2} \pi_{c_2}) +$$

$$\mu(\rho^2 \pi_{d_{1,1}} \pi_{d_{1,2}} + s^2 \pi_{c_1} \pi_{c_2}) +$$

$$\rho^2 s \pi_{d_{1,1}} \pi_{d_{1,2}} \pi_{c_1} + \rho s^2 \pi_{c_1} \pi_{c_1} \pi_{c_2} + s^3 \pi_{c_1} \pi_{c_2} \pi_{c_2} =$$

$$-\mu^3 +$$

$$-\mu^2(\rho\pi_{c_1} + s\pi_{c_2} + s) +$$

$$-\mu(\rho s \pi_{c_1} + s^2 \pi_{c_2})$$

Note that the terms of this equation are all parameters of the compound model, plus the within-group rate ρ . This arises because some of the terms with distinct model frequencies in the determinant calculation cancel out, and rearrangement and factorisation allows distinct frequencies to be replaced by compound model frequencies. The eigenvalues of \mathbf{Q}_D can thus be expressed in terms of the eigenvalues and parameters of \mathbf{Q}_C , as in Equation 4.11:

$$\mu_1 = \lambda_1 = -s, \mu_2 = q_{c_1 c_1} - \rho \pi_{c_1} = -(\rho \pi_{c_1} + s \pi_{c_2}), \mu_3 = \lambda_2 = 0$$

■

Appendix B

Example of calculations for the RY model

The RY model is a two-state model in nucleotides are classified as purines ('R') or pyrimidines ('Y'), and can be defined as a model with two compound states $c_1 = R$ and $c_2 = Y$, with associated distinct states $d_{1,1} = A, d_{1,2} = G, d_{2,1} = C$, and $d_{2,2} = T$. The rate matrices for the compound and distinct models are (from equations 4.1 and 4.2 in Chapter 4), respectively:

$$\mathbf{Q}_C = \begin{pmatrix} -s\pi_Y & s\pi_Y \\ s\pi_R & -s\pi_R \end{pmatrix}$$

and

$$\mathbf{Q}_D = \begin{pmatrix} -\rho_R\pi_G - s\pi_Y & \rho_R\pi_G & s\pi_C & s\pi_T \\ \rho_R\pi_A & -\rho_R\pi_A - s\pi_Y & s\pi_C & s\pi_T \\ s\pi_A & s\pi_G & -\rho_Y\pi_T - s\pi_R & \rho_Y\pi_T \\ s\pi_A & s\pi_G & \rho_Y\pi_C & -\rho_Y\pi_C - s\pi_R \end{pmatrix}$$

where $s = s_{RY}$ in order to eliminate redundant notation. The eigenvalues and eigenvectors of \mathbf{Q}_C are:

$$\lambda_1 = -s, \lambda_2 = 0 \text{ and } \mathbf{u}_1 = \begin{pmatrix} 1 \\ -\frac{\pi_R}{\pi_Y} \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and the corresponding \mathbf{V}_C values are:

$$\mathbf{v}_1 = \begin{pmatrix} \pi_Y \\ \pi_R \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} -\pi_Y \\ \pi_Y \end{pmatrix}$$

It is possible, albeit rather long-winded, to analytically calculate the eigenvalues of \mathbf{Q}_D . This is done by multiplying the matrix by $-\mu\mathbf{I}$, where \mathbf{I} is the identity matrix, and calculating the determinant to get a quartic equation in μ :

$$\begin{aligned} & \mu^4 + (2s + \rho_R\pi_R + \rho_Y\pi_Y)\mu^3 + \\ & (s^2(1 + \pi_R\pi_Y) + s\rho_R\pi_R(1 + \pi_R) + s\rho_Y\pi_Y(1 + \pi_Y) + \rho_R\rho_Y\pi_R\pi_Y)\mu^2 + \\ & (s^3\pi_R\pi_Y + s^2\rho_R\pi_R\pi_R + s^2\rho_Y\pi_Y\pi_Y + s\rho_R\rho_Y\pi_R\pi_Y)\mu \end{aligned}$$

Note that the terms of this equation are all parameters of the compound model, plus the within-group rates. This comes about because many of the terms with distinct model frequencies in the determinant calculation cancel out, and rearrangement and factorisation allows the sum of distinct frequencies to be replaced by the pertinent compound state frequency. The eigenvalues of \mathbf{Q}_D are:

$$\mu_1 = -s, \mu_2 = -s\pi_Y - \rho_R\pi_R, \mu_3 = 0, \mu_4 = -s\pi_R - \rho_Y\pi_Y$$

which corresponds to equation 4.11. The eigenvectors are:

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 1 \\ -\frac{\pi_R}{\pi_Y} \\ -\frac{\pi_R}{\pi_Y} \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} -\frac{\pi_G}{\pi_A} \\ 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{u}_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{u}_4 = \begin{pmatrix} 0 \\ 0 \\ -\frac{\pi_T}{\pi_C} \\ 1 \end{pmatrix}$$

and the corresponding \mathbf{V}_D values are:

$$\mathbf{v}_1 = \begin{pmatrix} \frac{\pi_Y \pi_A}{\pi_R} \\ -\frac{\pi_A}{\pi_R} \\ \pi_A \\ 0 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} \frac{\pi_Y \pi_G}{\pi_R} \\ 1 - \frac{\pi_G}{\pi_R} \\ \pi_G \\ 0 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} -\pi_C \\ 0 \\ \pi_C \\ -\frac{\pi_C}{\pi_Y} \end{pmatrix}, \mathbf{v}_4 = \begin{pmatrix} -\pi_T \\ 0 \\ \pi_T \\ 1 - \frac{\pi_T}{\pi_Y} \end{pmatrix}$$

These results can be used to calculate the matrix exponential of \mathbf{Q}_D , and thus a transition probability matrix, consistent with equation 4.18:

$$\mathbf{P}_D(t) = \begin{pmatrix} PRR(t) \frac{\pi_A}{\pi_R} + \frac{\pi_G}{\pi_R} e^{(q_{RR} - \rho_R \pi_R)t} & PRR(t) \frac{\pi_G}{\pi_R} - \frac{\pi_G}{\pi_R} e^{(q_{RR} - \rho_R \pi_R)t} & PRY(t) \frac{\pi_C}{\pi_Y} & PRY(t) \frac{\pi_T}{\pi_Y} \\ PRR(t) \frac{\pi_A}{\pi_R} - \frac{\pi_A}{\pi_R} e^{(q_{RR} - \rho_R \pi_R)t} & PRR(t) \frac{\pi_G}{\pi_R} + \frac{\pi_A}{\pi_R} e^{(q_{RR} - \rho_R \pi_R)t} & PRY(t) \frac{\pi_C}{\pi_Y} & PRY(t) \frac{\pi_T}{\pi_Y} \\ P_YR(t) \frac{\pi_A}{\pi_R} & P_YR(t) \frac{\pi_G}{\pi_R} & P_{YY}(t) \frac{\pi_C}{\pi_Y} + \frac{\pi_T}{\pi_Y} e^{(q_{YY} - \rho_Y \pi_Y)t} & P_{YY}(t) \frac{\pi_T}{\pi_Y} - \frac{\pi_T}{\pi_Y} e^{(q_{YY} - \rho_Y \pi_Y)t} \\ P_YR(t) \frac{\pi_A}{\pi_R} & P_YR(t) \frac{\pi_G}{\pi_R} & P_{YY}(t) \frac{\pi_C}{\pi_Y} - \frac{\pi_C}{\pi_Y} e^{(q_{YY} - \rho_Y \pi_Y)t} & P_{YY}(t) \frac{\pi_T}{\pi_Y} + \frac{\pi_C}{\pi_Y} e^{(q_{YY} - \rho_Y \pi_Y)t} \end{pmatrix}$$

If the ρ_i are considered to be ‘saturated’, and effectively infinite, a four-state equivalent of the RY model can be expressed, enabling direct comparison with standard nucleotide models:

$$\mathbf{P}_D(t) = \begin{pmatrix} PRR(t) \frac{\pi_A}{\pi_R} & PRR(t) \frac{\pi_G}{\pi_R} & PRY(t) \frac{\pi_C}{\pi_Y} & PRY(t) \frac{\pi_T}{\pi_Y} \\ PRR(t) \frac{\pi_A}{\pi_R} & PRR(t) \frac{\pi_G}{\pi_R} & PRY(t) \frac{\pi_C}{\pi_Y} & PRY(t) \frac{\pi_T}{\pi_Y} \\ P_YR(t) \frac{\pi_A}{\pi_R} & P_YR(t) \frac{\pi_G}{\pi_R} & P_{YY}(t) \frac{\pi_C}{\pi_Y} & P_{YY}(t) \frac{\pi_T}{\pi_Y} \\ P_YR(t) \frac{\pi_A}{\pi_R} & P_YR(t) \frac{\pi_G}{\pi_R} & P_{YY}(t) \frac{\pi_C}{\pi_Y} & P_{YY}(t) \frac{\pi_T}{\pi_Y} \end{pmatrix}$$

And the likelihood calculated with the two-state RY model can be corrected with the following, from equation 4.20:

$$L_D = L_C \prod_{i=1}^T \prod_{j=1}^l \frac{\pi_{d_{ij}}}{\pi_{c_{ij}}}$$

for an alignment of T taxa and of length l , where d_{ij} is the nucleotide in the i th taxon at the j th site, and c_{ij} is the purine/pyrimidine state.

Appendix C

Dinucleotide model definitions

The following model definitions are adapted from the PHASE manual (<http://tinyurl.com/phase-manual>) and Savill *et al.* (2001). Equilibrium frequencies are indicated by π_i and exchangeability parameters by α_j . The 16D, 16E and 16F models include two parameters describing the tendency toward stable base pairs (λ) and toward wobble base pairs (ϕ).

Table C.1: Model 7A

$$Q = \begin{pmatrix} \begin{array}{cccccc} AU & GU & GC & UA & UG & MM \\ * & \pi_{GU}\alpha_1 & \pi_{GC} & \pi_{UA}\alpha_2 & \pi_{UG}\alpha_3 & \pi_{MM}\alpha_5 \\ \pi_{AU}\alpha_1 & * & \pi_{GC}\alpha_6 & \pi_{UA}\alpha_7 & \pi_{UG}\alpha_8 & \pi_{MM}\alpha_{10} \\ \pi_{AU} & \pi_{GU}\alpha_6 & * & \pi_{UA}\alpha_{11} & \pi_{UG}\alpha_{12} & \pi_{MM}\alpha_{14} \\ \pi_{AU}\alpha_2 & \pi_{GU}\alpha_7 & \pi_{GC}\alpha_{11} & * & \pi_{UG}\alpha_{15} & \pi_{MM}\alpha_{17} \\ \pi_{AU}\alpha_3 & \pi_{GU}\alpha_8 & \pi_{GC}\alpha_{12} & \pi_{UA}\alpha_{15} & * & \pi_{MM}\alpha_{19} \\ \pi_{AU}\alpha_4 & \pi_{GU}\alpha_9 & \pi_{GC}\alpha_{13} & \pi_{UA}\alpha_{16} & \pi_{UG}\alpha_{18} & * \\ \pi_{AU}\alpha_5 & \pi_{GU}\alpha_{10} & \pi_{GC}\alpha_{14} & \pi_{UA}\alpha_{17} & \pi_{UG}\alpha_{19} & \pi_{CG}\alpha_{20} \\ * & & & & & * \end{array} \end{pmatrix}$$

Table C.2: Model 7B

$$Q = \begin{pmatrix} AU & GU & GC & UA & UG & CG & MM \\ * & \frac{\pi_{GU+UG}}{2} \alpha_1 & \frac{\pi_{GC+CG}}{2} & \frac{\pi_{AU+UA}}{2} \alpha_2 & \frac{\pi_{GU+UG}}{2} \alpha_3 & \frac{\pi_{GC+CG}}{2} \alpha_4 & \pi_{MM} \alpha_5 \\ \frac{\pi_{AU+UA}}{2} \alpha_1 & * & \frac{\pi_{GC+CG}}{2} \alpha_6 & \frac{\pi_{AU+UA}}{2} \alpha_7 & \frac{\pi_{GU+UG}}{2} \alpha_8 & \frac{\pi_{GC+CG}}{2} \alpha_9 & \pi_{MM} \alpha_{10} \\ \frac{\pi_{AU+UA}}{2} & \frac{\pi_{GU+UG}}{2} \alpha_6 & * & \frac{\pi_{AU+UA}}{2} \alpha_{11} & \frac{\pi_{GU+UG}}{2} \alpha_{12} & \frac{\pi_{GC+CG}}{2} \alpha_{13} & \pi_{MM} \alpha_{14} \\ \frac{\pi_{AU+UA}}{2} \alpha_2 & \frac{\pi_{GU+UG}}{2} \alpha_7 & \frac{\pi_{GC+CG}}{2} \alpha_{11} & * & \frac{\pi_{GU+UG}}{2} \alpha_{15} & \frac{\pi_{GC+CG}}{2} \alpha_{16} & \pi_{MM} \alpha_{17} \\ \frac{\pi_{AU+UA}}{2} \alpha_3 & \frac{\pi_{GU+UG}}{2} \alpha_8 & \frac{\pi_{GC+CG}}{2} \alpha_{12} & \frac{\pi_{AU+UA}}{2} \alpha_{15} & * & \frac{\pi_{GC+CG}}{2} \alpha_{18} & \pi_{MM} \alpha_{19} \\ \frac{\pi_{AU+UA}}{2} \alpha_4 & \frac{\pi_{GU+UG}}{2} \alpha_9 & \frac{\pi_{GC+CG}}{2} \alpha_{13} & \frac{\pi_{AU+UA}}{2} \alpha_{16} & \frac{\pi_{GU+UG}}{2} \alpha_{18} & * & \pi_{MM} \alpha_{20} \\ \frac{\pi_{AU+UA}}{2} \alpha_5 & \frac{\pi_{GU+UG}}{2} \alpha_{10} & \frac{\pi_{GC+CG}}{2} \alpha_{14} & \frac{\pi_{AU+UA}}{2} \alpha_{17} & \frac{\pi_{GU+UG}}{2} \alpha_{19} & \frac{\pi_{GC+CG}}{2} \alpha_{20} & * \end{pmatrix}$$

Table C.3: Model 7C

$$Q = \begin{pmatrix} & AU & GU & GC & UA & UG & CG & MM \\ AU & * & \pi_{GU} & 0 & 0 & 0 & 0 & \pi_{MM} \alpha_4 \\ GU & \pi_{AU} & * & \pi_{GC} \alpha_1 & 0 & 0 & 0 & \pi_{MM} \alpha_5 \\ GC & 0 & \pi_{GU} \alpha_1 & * & 0 & 0 & 0 & \pi_{MM} \alpha_6 \\ UA & 0 & 0 & 0 & * & \pi_{UG} \alpha_2 & 0 & \pi_{MM} \alpha_7 \\ UG & 0 & 0 & 0 & \pi_{UA} \alpha_2 & * & \pi_{CG} \alpha_3 & \pi_{MM} \alpha_8 \\ CG & 0 & 0 & 0 & 0 & \pi_{UG} \alpha_3 & * & \pi_{MM} \alpha_9 \\ MM & \pi_{AU} \alpha_4 & \pi_{GU} \alpha_5 & \pi_{GC} \alpha_6 & \pi_{UA} \alpha_7 & \pi_{UG} \alpha_8 & \pi_{CG} \alpha_9 & * \end{pmatrix}$$

Table C.4: Model 7D

$$Q = \begin{pmatrix} & AU & GU & GC & UA & UG & CG & MM \\ AU & * & \pi_{GU}\alpha_1 & \pi_{GC} & \pi_{UA}\alpha_2 & \pi_{UG}\alpha_2 & \pi_{CG}\alpha_2 & \pi_{MM}\alpha_3 \\ GU & \pi_{AU}\alpha_1 & * & \pi_{GC}\alpha_1 & \pi_{UA}\alpha_2 & \pi_{UG}\alpha_2 & \pi_{CG}\alpha_2 & \pi_{MM}\alpha_3 \\ GC & \pi_{AU} & \pi_{GU}\alpha_1 & * & \pi_{UA}\alpha_2 & \pi_{UG}\alpha_2 & \pi_{CG}\alpha_2 & \pi_{MM}\alpha_3 \\ UA & \pi_{AU}\alpha_2 & \pi_{GU}\alpha_2 & \pi_{GC}\alpha_2 & * & \pi_{UG}\alpha_1 & \pi_{CG} & \pi_{MM}\alpha_3 \\ UG & \pi_{AU}\alpha_2 & \pi_{GU}\alpha_2 & \pi_{GC}\alpha_2 & \pi_{UA}\alpha_1 & * & \pi_{CG}\alpha_1 & \pi_{MM}\alpha_3 \\ CG & \pi_{AU}\alpha_2 & \pi_{GU}\alpha_2 & \pi_{GC}\alpha_2 & \pi_{UA} & \pi_{UG}\alpha_1 & * & \pi_{MM}\alpha_3 \\ MM & \pi_{AU}\alpha_3 & \pi_{GU}\alpha_3 & \pi_{GC}\alpha_3 & \pi_{UA}\alpha_3 & \pi_{UG}\alpha_3 & \pi_{CG}\alpha_3 & * \end{pmatrix}$$

Table C.5: Model 7E

$$Q = \begin{pmatrix} & AU & GU & GC & UA & UG & CG & MM \\ AU & * & \pi_{GU} & 0 & 0 & 0 & 0 & \pi_{MM} \alpha_1 \\ GU & \pi_{AU} & * & \pi_{GC} & 0 & 0 & 0 & \pi_{MM} \alpha_1 \\ GC & 0 & \pi_{GU} & * & 0 & 0 & 0 & \pi_{MM} \alpha_1 \\ UA & 0 & 0 & 0 & * & \pi_{UG} & 0 & \pi_{MM} \alpha_1 \\ UG & 0 & 0 & 0 & \pi_{UA} & * & \pi_{CG} & \pi_{MM} \alpha_1 \\ CG & 0 & 0 & 0 & 0 & \pi_{UG} & * & \pi_{MM} \alpha_1 \\ MM & \pi_{AU} \alpha_1 & \pi_{GU} \alpha_1 & \pi_{GC} \alpha_1 & \pi_{UA} \alpha_1 & \pi_{UG} \alpha_1 & \pi_{CG} \alpha_1 & * \end{pmatrix}$$

Table C.6: Model 7F

$$Q = \begin{pmatrix} AU & GU & GC & UA & UG & CG & MM \\ AU & * & \frac{\pi_{GU+UG}}{2} \alpha_1 & \frac{\pi_{AU+UA}}{2} \alpha_2 & \frac{\pi_{GU+UG}}{2} \alpha_2 & \frac{\pi_{GC+CG}}{2} \alpha_2 & \pi_{MM} \alpha_3 \\ GU & \frac{\pi_{AU+UA}}{2} \alpha_1 & * & \frac{\pi_{AU+UA}}{2} \alpha_2 & \frac{\pi_{GU+UG}}{2} \alpha_2 & \frac{\pi_{GC+CG}}{2} \alpha_2 & \pi_{MM} \alpha_3 \\ GC & \frac{\pi_{AU+UA}}{2} \alpha_1 & \frac{\pi_{GU+UG}}{2} \alpha_1 & * & \frac{\pi_{GU+UG}}{2} \alpha_2 & \frac{\pi_{GC+CG}}{2} \alpha_2 & \pi_{MM} \alpha_3 \\ UA & \frac{\pi_{AU+UA}}{2} \alpha_2 & \frac{\pi_{GU+UG}}{2} \alpha_2 & * & \frac{\pi_{GU+UG}}{2} \alpha_1 & \frac{\pi_{GC+CG}}{2} \alpha_2 & \pi_{MM} \alpha_3 \\ UG & \frac{\pi_{AU+UA}}{2} \alpha_2 & \frac{\pi_{GU+UG}}{2} \alpha_2 & \frac{\pi_{AU+UA}}{2} \alpha_1 & * & \frac{\pi_{GC+CG}}{2} \alpha_1 & \pi_{MM} \alpha_3 \\ CG & \frac{\pi_{AU+UA}}{2} \alpha_2 & \frac{\pi_{GU+UG}}{2} \alpha_2 & \frac{\pi_{AU+UA}}{2} \alpha_2 & \frac{\pi_{GU+UG}}{2} \alpha_1 & * & \pi_{MM} \alpha_3 \\ MM & \frac{\pi_{AU+UA}}{2} \alpha_3 & \frac{\pi_{GU+UG}}{2} \alpha_3 & \frac{\pi_{AU+UA}}{2} \alpha_3 & \frac{\pi_{GU+UG}}{2} \alpha_3 & \frac{\pi_{GC+CG}}{2} \alpha_3 & * \end{pmatrix}$$

Table C.7: Model 7G

$$Q = \begin{pmatrix} AU & AU & GU & GC & UA & UG & CG & MM \\ * & \frac{\pi_{AU+UA}}{2} & \frac{\pi_{GU+UG}}{2} & 0 & 0 & 0 & 0 & \pi_{MM}\alpha \\ \frac{\pi_{AU+UA}}{2} & * & * & \frac{\pi_{GC+CG}}{2} & 0 & 0 & 0 & \pi_{MM}\alpha \\ 0 & \frac{\pi_{AU+UA}}{2} & \frac{\pi_{GU+UG}}{2} & * & 0 & 0 & 0 & \pi_{MM}\alpha \\ 0 & 0 & 0 & 0 & * & \frac{\pi_{GU+UG}}{2} & 0 & \pi_{MM}\alpha \\ 0 & 0 & 0 & 0 & \frac{\pi_{AU+UA}}{2} & * & \frac{\pi_{GC+CG}}{2} & \pi_{MM}\alpha \\ 0 & 0 & 0 & 0 & 0 & \frac{\pi_{GU+UG}}{2} & * & \pi_{MM}\alpha \\ \frac{\pi_{AU+UA}}{2} \alpha & \frac{\pi_{GU+UG}}{2} \alpha & \frac{\pi_{GC+CG}}{2} \alpha & \frac{\pi_{AU+UA}}{2} \alpha & \frac{\pi_{GU+UG}}{2} \alpha & \frac{\pi_{GC+CG}}{2} \alpha & * & * \end{pmatrix}$$

Table C.8: Model 16A

	AU	GU	GC	UA	UG	CG	AA	AG	AC	GA	GG	CA	CC	CU	UC	UU
AU	*	$\pi_{GU}\alpha_1$	π_{GC}	$\pi_{UA}\alpha_2$	$\pi_{UG}\alpha_2$	$\pi_{CG}\alpha_2$	$\pi_{AA}\alpha_3$	$\pi_{AG}\alpha_3$	$\pi_{AC}\alpha_3$	0	0	0	0	$\pi_{CU}\alpha_3$	0	$\pi_{UU}\alpha_3$
GU	$\pi_{AU}\alpha_1$	*	$\pi_{GC}\alpha_1$	$\pi_{UA}\alpha_2$	$\pi_{UG}\alpha_2$	$\pi_{CG}\alpha_2$	0	0	0	$\pi_{GA}\alpha_3$	$\pi_{GG}\alpha_3$	0	0	$\pi_{CU}\alpha_3$	0	$\pi_{UU}\alpha_3$
GC	$\pi_{AU}\alpha_2$	$\pi_{GU}\alpha_1$	*	$\pi_{UA}\alpha_2$	$\pi_{UG}\alpha_2$	$\pi_{CG}\alpha_2$	0	0	$\pi_{AC}\alpha_3$	$\pi_{GA}\alpha_3$	$\pi_{GG}\alpha_3$	0	$\pi_{CC}\alpha_3$	0	$\pi_{UC}\alpha_3$	0
UA	$\pi_{AU}\alpha_2$	$\pi_{GU}\alpha_2$	$\pi_{GC}\alpha_2$	*	$\pi_{UG}\alpha_1$	π_{CG}	$\pi_{AA}\alpha_3$	0	0	$\pi_{GA}\alpha_3$	0	$\pi_{CA}\alpha_3$	0	0	$\pi_{UC}\alpha_3$	$\pi_{UU}\alpha_3$
UG	$\pi_{AU}\alpha_2$	$\pi_{GU}\alpha_2$	$\pi_{GC}\alpha_2$	$\pi_{UA}\alpha_1$	*	$\pi_{CG}\alpha_1$	0	$\pi_{AG}\alpha_3$	0	0	$\pi_{GG}\alpha_3$	0	0	0	$\pi_{UC}\alpha_3$	$\pi_{UU}\alpha_3$
CG	$\pi_{AU}\alpha_2$	$\pi_{GU}\alpha_2$	$\pi_{GC}\alpha_2$	π_{UA}	$\pi_{UG}\alpha_1$	*	0	$\pi_{AG}\alpha_3$	0	0	$\pi_{GG}\alpha_3$	$\pi_{CA}\alpha_3$	$\pi_{CC}\alpha_3$	$\pi_{CU}\alpha_3$	0	0
AA	$\pi_{AU}\alpha_3$	0	0	$\pi_{UA}\alpha_3$	0	0	*	$\pi_{AG}\alpha_4$	$\pi_{AC}\alpha_4$	$\pi_{GA}\alpha_4$	0	$\pi_{CA}\alpha_4$	0	0	0	0
AG	$\pi_{AU}\alpha_3$	0	0	0	$\pi_{UG}\alpha_3$	$\pi_{CG}\alpha_3$	$\pi_{AA}\alpha_4$	*	$\pi_{AC}\alpha_4$	0	$\pi_{GG}\alpha_4$	0	0	0	0	0
AC	$\pi_{AU}\alpha_3$	0	$\pi_{GC}\alpha_3$	0	0	0	$\pi_{AA}\alpha_4$	$\pi_{AG}\alpha_4$	*	0	0	0	$\pi_{CC}\alpha_4$	0	$\pi_{UC}\alpha_4$	0
GA	0	$\pi_{GU}\alpha_3$	$\pi_{GC}\alpha_3$	$\pi_{UA}\alpha_3$	0	0	$\pi_{AA}\alpha_4$	0	0	*	$\pi_{GG}\alpha_4$	$\pi_{CA}\alpha_4$	0	0	0	0
GG	0	$\pi_{GU}\alpha_3$	$\pi_{GC}\alpha_3$	0	$\pi_{UG}\alpha_3$	$\pi_{CG}\alpha_3$	0	$\pi_{AG}\alpha_4$	0	$\pi_{GA}\alpha_4$	*	0	0	0	0	0
CA	0	0	0	$\pi_{UA}\alpha_3$	0	$\pi_{CG}\alpha_3$	$\pi_{AA}\alpha_4$	0	0	$\pi_{GA}\alpha_4$	0	*	$\pi_{CC}\alpha_4$	$\pi_{CU}\alpha_4$	0	0
CC	0	0	$\pi_{GC}\alpha_3$	0	0	$\pi_{CG}\alpha_3$	0	0	$\pi_{AC}\alpha_4$	0	0	$\pi_{CA}\alpha_4$	*	$\pi_{UC}\alpha_4$	0	0
CU	$\pi_{AU}\alpha_3$	$\pi_{GU}\alpha_3$	0	0	0	$\pi_{CG}\alpha_3$	0	0	0	0	0	$\pi_{CA}\alpha_4$	$\pi_{CC}\alpha_4$	*	0	$\pi_{UU}\alpha_4$
UC	0	0	$\pi_{GC}\alpha_3$	$\pi_{UA}\alpha_3$	$\pi_{UG}\alpha_3$	0	0	0	$\pi_{AC}\alpha_4$	0	0	0	$\pi_{CC}\alpha_4$	0	*	$\pi_{UU}\alpha_4$
UU	$\pi_{AU}\alpha_3$	$\pi_{GU}\alpha_3$	0	$\pi_{UA}\alpha_3$	$\pi_{UG}\alpha_3$	0	0	0	0	0	0	0	0	$\pi_{CU}\alpha_4$	$\pi_{UC}\alpha_4$	*

Q =

Table C.9: Model 16B

$$Q = \begin{pmatrix}
\begin{array}{cccccccccccccccc}
AU & GU & GC & UA & UG & CG & AA & AG & AC & GA & GG & CA & CC & CU & UC & UU \\
* & \pi_{GU} & 0 & 0 & 0 & 0 & \pi_{AA} & \pi_{AG} & \pi_{AC} & 0 & 0 & 0 & 0 & \pi_{CU} & 0 & \pi_{UU} \\
\pi_{AU} & * & \pi_{GC} & 0 & 0 & 0 & 0 & 0 & 0 & \pi_{GA} & \pi_{GG} & 0 & 0 & \pi_{CU} & 0 & \pi_{UU} \\
0 & \pi_{GU} & * & 0 & 0 & 0 & 0 & 0 & \pi_{AC} & \pi_{GA} & \pi_{GG} & 0 & \pi_{CC} & 0 & \pi_{UC} & 0 \\
0 & 0 & 0 & * & \pi_{UG} & 0 & \pi_{AA} & 0 & 0 & \pi_{GA} & 0 & \pi_{CA} & 0 & 0 & \pi_{UC} & \pi_{UU} \\
0 & 0 & 0 & \pi_{UA} & * & \pi_{CG} & 0 & \pi_{AG} & 0 & 0 & \pi_{GG} & 0 & 0 & 0 & \pi_{UC} & \pi_{UU} \\
0 & 0 & 0 & 0 & \pi_{UG} & * & 0 & \pi_{AG} & 0 & 0 & \pi_{GG} & \pi_{CA} & \pi_{CC} & \pi_{CU} & 0 & 0 \\
\pi_{AU} & 0 & 0 & \pi_{UA} & 0 & 0 & * & \pi_{AG} & \pi_{AC} & \pi_{GA} & 0 & \pi_{CA} & 0 & 0 & 0 & 0 \\
\pi_{AU} & 0 & 0 & 0 & \pi_{UG} & \pi_{CG} & \pi_{AA} & * & \pi_{AC} & 0 & \pi_{GG} & 0 & 0 & 0 & 0 & 0 \\
\pi_{AU} & 0 & \pi_{GC} & 0 & 0 & 0 & \pi_{AA} & \pi_{AG} & * & 0 & 0 & 0 & \pi_{CC} & 0 & \pi_{UC} & 0 \\
0 & \pi_{GU} & \pi_{GC} & \pi_{UA} & 0 & 0 & \pi_{AA} & 0 & 0 & * & \pi_{GG} & \pi_{CA} & 0 & 0 & 0 & 0 \\
0 & \pi_{GU} & \pi_{GC} & 0 & \pi_{UG} & \pi_{CG} & 0 & \pi_{AG} & 0 & \pi_{GA} & * & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \pi_{UA} & 0 & \pi_{CG} & \pi_{AA} & 0 & 0 & \pi_{GA} & 0 & * & \pi_{CC} & \pi_{CU} & 0 & 0 \\
0 & 0 & \pi_{GC} & 0 & 0 & \pi_{CG} & 0 & 0 & \pi_{AC} & 0 & 0 & \pi_{CA} & * & \pi_{CU} & \pi_{UC} & 0 \\
\pi_{AU} & \pi_{GU} & 0 & 0 & 0 & \pi_{CG} & 0 & 0 & 0 & 0 & 0 & \pi_{CA} & \pi_{CC} & * & 0 & \pi_{UU} \\
0 & 0 & \pi_{GC} & \pi_{UA} & \pi_{UG} & 0 & 0 & 0 & \pi_{AC} & 0 & 0 & \pi_{CA} & \pi_{CC} & 0 & * & \pi_{UU} \\
\pi_{AU} & \pi_{GU} & 0 & \pi_{UA} & \pi_{UG} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \pi_{CU} & \pi_{UC} & *
\end{array}
\end{pmatrix}$$

Table C.10: Model 16C

$$Q = \begin{pmatrix} AU & GU & GC & UA & UG & CG & AA & AG & AC & GA & GG & CA & CC & CU & UC & UU \\ AU & * & \pi_{GU} & 0 & 0 & 0 & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & 0 & 0 & 0 & 0 & \frac{\pi_{MM}}{10} & 0 & \frac{\pi_{MM}}{10} \\ GU & \pi_{AU} & * & \pi_{GC} & 0 & 0 & 0 & 0 & 0 & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & 0 & 0 & \frac{\pi_{MM}}{10} & 0 & \frac{\pi_{MM}}{10} \\ GC & 0 & \pi_{GU} & * & 0 & 0 & 0 & 0 & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & 0 & \frac{\pi_{MM}}{10} & 0 & \frac{\pi_{MM}}{10} & 0 \\ UA & 0 & 0 & 0 & * & \pi_{UG} & 0 & 0 & 0 & \frac{\pi_{MM}}{10} & 0 & \frac{\pi_{MM}}{10} & 0 & 0 & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} \\ UG & 0 & 0 & 0 & \pi_{UA} & * & \pi_{CG} & 0 & 0 & \frac{\pi_{MM}}{10} & 0 & 0 & 0 & 0 & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} \\ CG & 0 & 0 & 0 & 0 & \pi_{UG} & * & 0 & 0 & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & 0 & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & 0 & 0 \\ AA & \pi_{AU} & 0 & 0 & \pi_{UA} & 0 & * & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & 0 & \frac{\pi_{MM}}{10} & 0 & 0 & 0 & 0 \\ AG & \pi_{AU} & 0 & 0 & 0 & \pi_{UG} & \pi_{CG} & * & \frac{\pi_{MM}}{10} & 0 & \frac{\pi_{MM}}{10} & 0 & 0 & 0 & 0 & 0 \\ AC & \pi_{AU} & 0 & \pi_{GC} & 0 & 0 & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & * & 0 & 0 & 0 & \frac{\pi_{MM}}{10} & 0 & \frac{\pi_{MM}}{10} & 0 \\ GA & 0 & \pi_{GU} & \pi_{GC} & \pi_{UA} & 0 & 0 & 0 & 0 & * & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & 0 & 0 & 0 & 0 \\ GG & 0 & \pi_{GU} & \pi_{GC} & 0 & \pi_{UG} & \pi_{CG} & 0 & 0 & \frac{\pi_{MM}}{10} & * & 0 & 0 & 0 & 0 & 0 \\ CA & 0 & 0 & 0 & \pi_{UA} & 0 & \pi_{CG} & 0 & 0 & \frac{\pi_{MM}}{10} & 0 & * & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & 0 & 0 \\ CC & 0 & 0 & \pi_{GC} & 0 & 0 & \pi_{CG} & 0 & \frac{\pi_{MM}}{10} & 0 & 0 & \frac{\pi_{MM}}{10} & * & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & 0 \\ CU & \pi_{AU} & \pi_{GU} & 0 & 0 & 0 & \pi_{CG} & 0 & 0 & 0 & 0 & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & * & 0 & \frac{\pi_{MM}}{10} \\ UC & 0 & 0 & \pi_{GC} & \pi_{UA} & \pi_{UG} & 0 & 0 & \frac{\pi_{MM}}{10} & 0 & 0 & 0 & \frac{\pi_{MM}}{10} & 0 & * & \frac{\pi_{MM}}{10} \\ UU & \pi_{AU} & \pi_{GU} & 0 & \pi_{UA} & \pi_{UG} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\pi_{MM}}{10} & \frac{\pi_{MM}}{10} & * \end{pmatrix}$$

Table C.11: Model 16D

	AU	GU	GC	UA	UG	CG	AA	AG	AC	GA	GG	CA	CC	CU	UC	UU
AU	*	$\frac{\phi\pi_G}{\lambda}$	0	0	0	0	$\frac{\alpha\pi_A}{\lambda}$	$\frac{\alpha\pi_G}{\lambda}$	$\frac{\pi_C}{\lambda}$	0	0	0	0	$\frac{\alpha\pi_C}{\lambda}$	0	$\frac{\alpha\pi_U}{\lambda}$
GU	$\frac{\lambda\pi_A}{\phi}$	*	$\frac{\lambda\pi_C}{\phi}$	0	0	0	0	0	0	$\frac{\alpha\pi_A}{\phi}$	$\frac{\alpha\pi_G}{\phi}$	0	0	$\frac{\alpha\pi_C}{\phi}$	0	$\frac{\alpha\pi_U}{\phi}$
GC	0	$\frac{\phi\pi_U}{\lambda}$	*	0	0	0	0	0	$\frac{\pi_A}{\lambda}$	$\frac{\alpha\pi_A}{\lambda}$	$\frac{\alpha\pi_G}{\lambda}$	0	$\frac{\alpha\pi_C}{\lambda}$	0	$\frac{\alpha\pi_U}{\lambda}$	0
UA	0	0	0	*	$\frac{\phi\pi_G}{\lambda}$	0	$\frac{\alpha\pi_A}{\lambda}$	0	0	$\frac{\alpha\pi_G}{\lambda}$	0	$\frac{\alpha\pi_C}{\lambda_1}$	0	0	$\frac{\alpha\pi_C}{\lambda}$	$\frac{\alpha\pi_U}{\lambda}$
UG	0	0	0	$\frac{\lambda\pi_A}{\phi}$	*	$\frac{\lambda\pi_C}{\phi}$	0	$\frac{\alpha\pi_A}{\phi}$	0	0	$\frac{\alpha\pi_G}{\phi}$	0	0	0	$\frac{\alpha\pi_C}{\phi}$	$\frac{\alpha\pi_U}{\phi}$
CG	0	0	0	0	$\frac{\phi\pi_U}{\lambda}$	*	0	$\frac{\alpha\pi_A}{\lambda}$	0	0	$\frac{\alpha\pi_G}{\lambda}$	$\frac{\pi_A}{\lambda}$	$\frac{\alpha\pi_C}{\lambda}$	$\frac{\alpha\pi_U}{\lambda}$	0	0
AA	$\frac{\alpha\pi_U}{\lambda}$	0	0	$\alpha\lambda\pi_A$	0	0	*	π_G	$\alpha\pi_C$	π_G	0	π_C	0	0	0	0
AG	$\frac{\alpha\pi_U}{\lambda}$	0	0	0	$\alpha\phi\pi_G$	$\alpha\lambda\pi_C$	π_A	*	$\alpha\pi_C$	0	π_G	0	0	0	0	0
AC	$\frac{\pi_U}{\lambda}$	0	$\lambda\pi_G$	0	0	0	$\alpha\pi_A$	$\alpha\pi_G$	*	0	0	0	$\alpha\pi_{CC}$	0	$\alpha\pi_U$	0
GA	0	$\alpha\phi\pi_U$	$\alpha\lambda\pi_C$	$\alpha\lambda\pi_A$	0	0	$\alpha\pi_A$	0	0	*	π_G	$\alpha\pi_C$	0	0	0	0
GG	0	$\alpha\phi\pi_U$	$\alpha\lambda\pi_C$	0	$\alpha\phi\pi_G$	$\alpha\lambda\pi_C$	0	π_A	0	π_A	*	0	0	0	0	0
CA	0	0	0	$\alpha\lambda\pi_A$	0	$\lambda\pi_G$	π_A	0	0	$\alpha\pi_G$	0	*	$\alpha\pi_C$	$\alpha\pi_U$	0	0
CC	0	0	$\alpha\lambda\pi_G$	0	0	$\alpha\lambda\pi_G$	0	0	$\alpha\pi_A$	0	0	$\alpha\pi_A$	*	π_U	π_U	0
CU	$\frac{\alpha\pi_A}{\lambda}$	$\alpha\phi\pi_G$	0	0	0	$\alpha\lambda\pi_G$	0	0	0	0	0	$\alpha\pi_A$	π_C	*	0	π_U
UC	0	0	$\alpha\lambda\pi_G$	$\alpha\lambda\pi_A$	$\alpha\phi\pi_G$	0	0	0	$\alpha\pi_A$	0	0	0	π_C	0	*	π_U
UU	$\frac{\alpha\pi_A}{\lambda}$	$\alpha\phi\pi_G$	0	$\alpha\lambda\pi_A$	$\alpha\phi\pi_G$	0	0	0	0	0	0	0	0	π_C	π_C	*

$Q =$

Table C.12: Model 16E

	AU	GU	GC	UA	UG	CG	AA	AG	AC	GA	GG	CA	CC	CU	UC	UU
AU	*	$\frac{\pi_G}{\lambda}$	0	0	0	0	$\frac{\alpha\pi_A}{\lambda}$	$\frac{\alpha\pi_G}{\lambda}$	$\frac{\pi_C}{\lambda}$	0	0	0	0	$\frac{\alpha\pi_C}{\lambda}$	0	$\frac{\alpha\pi_U}{\lambda}$
GU	$\lambda\pi_A$	*	$\lambda\pi_C$	0	0	0	0	0	0	$\alpha\pi_A$	$\alpha\pi_G$	0	0	$\alpha\pi_C$	0	$\alpha\pi_U$
GC	0	$\frac{\pi_U}{\lambda}$	*	0	0	0	0	0	$\frac{\pi_A}{\lambda}$	$\frac{\alpha\pi_A}{\lambda}$	$\frac{\alpha\pi_G}{\lambda}$	0	$\frac{\alpha\pi_C}{\lambda}$	0	$\frac{\alpha\pi_U}{\lambda}$	0
UA	0	0	0	*	$\frac{\pi_G}{\lambda}$	0	$\frac{\alpha\pi_A}{\lambda}$	0	0	$\frac{\alpha\pi_G}{\lambda}$	0	$\frac{\alpha\pi_C}{\lambda_1}$	0	0	$\frac{\alpha\pi_U}{\lambda}$	0
UG	0	0	0	$\lambda\pi_A$	*	$\lambda\pi_C$	0	$\alpha\pi_A$	0	0	$\alpha\pi_G$	0	0	0	$\alpha\pi_C$	$\alpha\pi_U$
CG	0	0	0	0	$\frac{\pi_U}{\lambda}$	*	0	$\frac{\alpha\pi_A}{\lambda}$	0	0	$\frac{\alpha\pi_G}{\lambda}$	$\frac{\pi_A}{\lambda}$	$\frac{\alpha\pi_C}{\lambda}$	$\frac{\alpha\pi_U}{\lambda}$	0	0
AA	$\frac{\alpha\pi_U}{\lambda}$	0	0	$\alpha\lambda\pi_A$	0	0	*	π_G	$\alpha\pi_C$	π_G	0	π_C	0	0	0	0
AG	$\frac{\alpha\pi_U}{\lambda}$	0	0	0	$\alpha\pi_G$	$\alpha\lambda\pi_C$	π_A	*	$\alpha\pi_C$	0	π_G	0	0	0	0	0
AC	$\frac{\pi_U}{\lambda}$	0	$\lambda\pi_G$	0	0	0	$\alpha\pi_A$	$\alpha\pi_G$	*	0	0	0	$\alpha\pi_{CC}$	0	$\alpha\pi_U$	0
GA	0	$\alpha\pi_U$	$\alpha\lambda\pi_C$	$\alpha\lambda\pi_A$	0	0	$\alpha\pi_A$	0	0	*	π_G	$\alpha\pi_C$	0	0	0	0
GG	0	$\alpha\pi_U$	$\alpha\lambda\pi_C$	0	$\alpha\pi_G$	$\alpha\lambda\pi_C$	0	π_A	0	π_A	*	0	0	0	0	0
CA	0	0	0	$\alpha\lambda\pi_A$	0	$\lambda\pi_G$	π_A	0	0	$\alpha\pi_G$	0	*	$\alpha\pi_C$	$\alpha\pi_U$	0	0
CC	0	0	$\alpha\lambda\pi_G$	0	0	$\alpha\lambda\pi_G$	0	0	$\alpha\pi_A$	0	0	$\alpha\pi_A$	*	π_U	π_U	0
CU	$\frac{\alpha\pi_A}{\lambda}$	$\alpha\pi_G$	0	0	0	$\alpha\lambda\pi_G$	0	0	0	0	0	$\alpha\pi_A$	π_C	*	0	π_U
UC	0	0	$\alpha\lambda\pi_G$	$\alpha\lambda\pi_A$	$\alpha\pi_G$	0	0	0	$\alpha\pi_A$	0	0	0	π_C	0	*	π_U
UU	$\frac{\alpha\pi_A}{\lambda}$	$\alpha\pi_G$	0	$\alpha\lambda\pi_A$	$\alpha\pi_G$	0	0	0	0	0	0	0	0	π_C	π_C	*

Q =

Table C.13: Model 16F

	AU	GU	GC	UA	UG	CG	AA	AG	AC	GA	GG	CA	CC	CU	UC	UU
AU	*	π_G	0	0	0	0	$\frac{\alpha\pi_A}{\lambda}$	$\frac{\alpha\pi_G}{\lambda}$	$\frac{\pi_C}{\lambda}$	0	0	0	0	$\frac{\alpha\pi_C}{\lambda}$	0	$\frac{\alpha\pi_U}{\lambda}$
GU	π_A	*	π_C	0	0	0	0	0	0	$\frac{\alpha\pi_A}{\lambda}$	$\frac{\alpha\pi_G}{\lambda}$	0	0	$\frac{\alpha\pi_C}{\lambda}$	0	$\frac{\alpha\pi_U}{\lambda}$
GC	0	π_U	*	0	0	0	0	0	$\frac{\pi_A}{\lambda}$	$\frac{\alpha\pi_A}{\lambda}$	$\frac{\alpha\pi_G}{\lambda}$	0	$\frac{\alpha\pi_C}{\lambda}$	0	$\frac{\alpha\pi_U}{\lambda}$	0
UA	0	0	0	*	π_G	0	$\frac{\alpha\pi_A}{\lambda}$	0	0	$\frac{\alpha\pi_G}{\lambda}$	0	$\frac{\alpha\pi_C}{\lambda_1}$	0	0	$\frac{\alpha\pi_C}{\lambda}$	$\frac{\alpha\pi_U}{\lambda}$
UG	0	0	0	π_A	*	π_C	0	$\frac{\alpha\pi_A}{\lambda}$	0	0	$\frac{\alpha\pi_G}{\lambda}$	0	0	0	$\frac{\alpha\pi_C}{\lambda}$	$\frac{\alpha\pi_U}{\lambda}$
CG	0	0	0	0	π_U	*	0	$\frac{\alpha\pi_A}{\lambda}$	0	0	$\frac{\alpha\pi_G}{\lambda}$	$\frac{\pi_A}{\lambda}$	$\frac{\alpha\pi_C}{\lambda}$	$\frac{\alpha\pi_U}{\lambda}$	0	0
AA	$\frac{\alpha\pi_U}{\lambda}$	0	0	$\alpha\lambda\pi_A$	0	0	*	π_G	$\alpha\pi_C$	π_G	0	π_C	0	0	0	0
AG	$\frac{\alpha\pi_U}{\lambda}$	0	0	0	$\alpha\lambda\pi_G$	$\alpha\lambda\pi_C$	π_A	*	$\alpha\pi_C$	0	π_G	0	0	0	0	0
AC	$\frac{\pi_U}{\lambda}$	0	$\lambda\pi_G$	0	0	0	$\alpha\pi_A$	$\alpha\pi_G$	*	0	0	0	$\alpha\pi_{CC}$	0	$\alpha\pi_U$	0
GA	0	$\alpha\lambda\pi_U$	$\alpha\lambda\pi_C$	$\alpha\lambda\pi_A$	0	0	$\alpha\pi_A$	0	0	*	π_G	$\alpha\pi_C$	0	0	0	0
GG	0	$\alpha\lambda\pi_U$	$\alpha\lambda\pi_C$	0	$\alpha\lambda\pi_G$	$\alpha\lambda\pi_C$	0	π_A	0	π_A	*	0	0	0	0	0
CA	0	0	0	$\alpha\lambda\pi_A$	0	$\lambda\pi_G$	π_A	0	0	$\alpha\pi_G$	0	*	$\alpha\pi_C$	$\alpha\pi_U$	0	0
CC	0	0	$\alpha\lambda\pi_G$	0	0	$\alpha\lambda\pi_G$	0	0	$\alpha\pi_A$	0	0	$\alpha\pi_A$	*	π_U	π_U	0
CU	$\frac{\alpha\pi_A}{\lambda}$	$\alpha\lambda\pi_G$	0	0	0	$\alpha\lambda\pi_G$	0	0	0	0	0	$\alpha\pi_A$	π_C	*	0	π_U
UC	0	0	$\alpha\lambda\pi_G$	$\alpha\lambda\pi_A$	$\alpha\lambda\pi_G$	0	0	0	$\alpha\pi_A$	0	0	0	π_C	0	*	π_U
UU	$\frac{\alpha\pi_A}{\lambda}$	$\alpha\lambda\pi_G$	0	$\alpha\lambda\pi_A$	$\alpha\lambda\pi_G$	0	0	0	0	0	0	0	0	π_C	π_C	*

Q =

Table C.14: Model 16I

	AU	GU	GC	UA	UG	CG	AA	AG	AC	GA	GG	CA	CC	CU	UC	UU
AU	*	π_{GU}	0	0	0	0	$\pi_{AA}\alpha_2$	$\pi_{AG}\alpha_5$	$\pi_{AC}\alpha_4$	0	0	0	0	$\pi_{CU}\alpha_1$	0	$\pi_{UU}\alpha_2$
GU	π_{AU}	*	$\pi_{GC}\alpha_4$	0	0	0	0	0	0	$\pi_{GA}\alpha_2$	$\pi_{GG}\alpha_5$	0	0	$\pi_{CU}\alpha_3$	0	$\pi_{UU}\alpha_5$
GC	0	$\pi_{GU}\alpha_4$	*	0	0	0	0	0	π_{AC}	$\pi_{GA}\alpha_1$	$\pi_{GG}\alpha_3$	0	$\pi_{CC}\alpha_3$	0	$\pi_{UC}\alpha_5$	0
UA	0	0	0	*	π_{UG}	0	$\pi_{AA}\alpha_2$	0	0	$\pi_{GA}\alpha_5$	0	$\pi_{CA}\alpha_4$	0	0	$\pi_{UC}\alpha_1$	$\pi_{UU}\alpha_2$
UG	0	0	0	π_{UA}	*	$\pi_{CG}\alpha_4$	0	$\pi_{AG}\alpha_2$	0	0	$\pi_{GG}\alpha_4$	0	0	0	$\pi_{UC}\alpha_3$	$\pi_{UU}\alpha_5$
CG	0	0	0	0	$\pi_{UG}\alpha_4$	*	0	$\pi_{AG}\alpha_1$	0	0	$\pi_{GG}\alpha_3$	π_{CA}	$\pi_{CC}\alpha_3$	$\pi_{CU}\alpha_5$	0	0
AA	$\pi_{AU}\alpha_2$	0	0	$\pi_{UA}\alpha_2$	0	0	*	π_{AG}	$\pi_{AC}\alpha_1$	π_{GA}	0	$\pi_{CA}\alpha_1$	0	0	0	0
AG	$\pi_{AU}\alpha_5$	0	0	0	$\pi_{UG}\alpha_2$	$\pi_{CG}\alpha_1$	π_{AA}	*	$\pi_{AC}\alpha_3$	0	π_{GG}	0	0	0	0	0
AC	$\pi_{AU}\alpha_4$	0	π_{GC}	0	0	0	$\pi_{AA}\alpha_1$	$\pi_{AG}\alpha_3$	*	0	0	0	$\pi_{CC}\alpha_1$	0	$\pi_{UC}\alpha_2$	0
GA	0	$\pi_{GU}\alpha_2$	$\pi_{GC}\alpha_1$	$\pi_{UA}\alpha_5$	0	0	π_{AA}	0	0	*	π_{GG}	$\pi_{CA}\alpha_3$	0	0	0	0
GG	0	$\pi_{GU}\alpha_5$	$\pi_{GC}\alpha_3$	0	$\pi_{UG}\alpha_4$	$\pi_{CG}\alpha_3$	0	π_{AG}	0	π_{GA}	*	0	0	0	0	0
CA	0	0	0	$\pi_{UA}\alpha_4$	0	π_{CG}	$\pi_{AA}\alpha_1$	0	0	$\pi_{GA}\alpha_3$	0	*	$\pi_{CC}\alpha_1$	$\pi_{CU}\alpha_2$	0	0
CC	0	0	$\pi_{GC}\alpha_3$	0	0	$\pi_{CG}\alpha_3$	0	0	$\pi_{AC}\alpha_1$	0	0	$\pi_{CA}\alpha_1$	*	$\pi_{UC}\alpha_4$	0	0
CU	$\pi_{AU}\alpha_1$	$\pi_{GU}\alpha_3$	0	0	0	$\pi_{CG}\alpha_5$	0	0	0	0	0	$\pi_{CA}\alpha_2$	$\pi_{CC}\alpha_4$	*	0	$\pi_{UU}\alpha_4$
UC	0	0	$\pi_{GC}\alpha_5$	$\pi_{UA}\alpha_1$	$\pi_{UG}\alpha_3$	0	0	0	$\pi_{AC}\alpha_2$	0	0	0	$\pi_{CC}\alpha_4$	0	*	$\pi_{UU}\alpha_4$
UU	$\pi_{AU}\alpha_2$	$\pi_{GU}\alpha_5$	0	$\pi_{UA}\alpha_2$	$\pi_{UG}\alpha_5$	0	0	0	0	0	0	0	0	$\pi_{CU}\alpha_4$	$\pi_{UC}\alpha_4$	*

$Q =$

Table C.15: Model 16J

	AU	GU	GC	UA	UG	CG	AA	AG	AC	GA	GG	CA	CC	CU	UC	UU
AU	*	π_{GU}	0	0	0	0	$\pi_{AA}\alpha_1$	$\pi_{AG}\alpha_1$	$\pi_{AC}\alpha_2$	0	0	0	0	$\pi_{CU}\alpha_1$	0	$\pi_{UU}\alpha_1$
GU	π_{AU}	*	$\pi_{GC}\alpha_2$	0	0	0	0	0	0	$\pi_{GA}\alpha_1$	$\pi_{GG}\alpha_1$	0	0	$\pi_{CU}\alpha_1$	0	$\pi_{UU}\alpha_1$
GC	0	$\pi_{GU}\alpha_2$	*	0	0	0	0	0	π_{AC}	$\pi_{GA}\alpha_1$	$\pi_{GG}\alpha_1$	0	$\pi_{CC}\alpha_1$	0	$\pi_{UC}\alpha_1$	0
UA	0	0	0	*	π_{UG}	0	$\pi_{AA}\alpha_1$	0	0	$\pi_{GA}\alpha_1$	0	$\pi_{CA}\alpha_2$	0	0	$\pi_{UC}\alpha_1$	$\pi_{UU}\alpha_1$
UG	0	0	0	π_{UA}	*	$\pi_{CG}\alpha_2$	0	$\pi_{AG}\alpha_1$	0	0	$\pi_{GG}\alpha_2$	0	0	0	$\pi_{UC}\alpha_1$	$\pi_{UU}\alpha_1$
CG	0	0	0	0	$\pi_{UG}\alpha_2$	*	0	$\pi_{AG}\alpha_1$	0	0	$\pi_{GG}\alpha_1$	π_{CA}	$\pi_{CC}\alpha_1$	$\pi_{CU}\alpha_1$	0	0
AA	$\pi_{AU}\alpha_1$	0	0	$\pi_{UA}\alpha_1$	0	0	*	π_{AG}	$\pi_{AC}\alpha_1$	π_{GA}	0	$\pi_{CA}\alpha_1$	0	0	0	0
AG	$\pi_{AU}\alpha_1$	0	0	0	$\pi_{UG}\alpha_1$	$\pi_{CG}\alpha_1$	π_{AA}	*	$\pi_{AC}\alpha_1$	0	π_{GG}	0	0	0	0	0
AC	$\pi_{AU}\alpha_2$	0	π_{GC}	0	0	0	$\pi_{AA}\alpha_1$	$\pi_{AG}\alpha_1$	*	0	0	0	$\pi_{CC}\alpha_1$	0	$\pi_{UC}\alpha_1$	0
GA	0	$\pi_{GU}\alpha_1$	$\pi_{GC}\alpha_1$	$\pi_{UA}\alpha_1$	0	0	π_{AA}	0	0	*	π_{GG}	$\pi_{CA}\alpha_1$	0	0	0	0
GG	0	$\pi_{GU}\alpha_1$	$\pi_{GC}\alpha_1$	0	$\pi_{UG}\alpha_2$	$\pi_{CG}\alpha_1$	0	π_{AG}	0	π_{GA}	*	0	0	0	0	0
CA	0	0	0	$\pi_{UA}\alpha_2$	0	π_{CG}	$\pi_{AA}\alpha_1$	0	0	$\pi_{GA}\alpha_1$	0	*	$\pi_{CC}\alpha_1$	$\pi_{CU}\alpha_1$	0	0
CC	0	0	$\pi_{GC}\alpha_1$	0	0	$\pi_{CG}\alpha_1$	0	0	$\pi_{AC}\alpha_1$	0	0	$\pi_{CA}\alpha_1$	*	$\pi_{CC}\alpha_1$	$\pi_{UC}\alpha_2$	0
CU	$\pi_{AU}\alpha_1$	$\pi_{GU}\alpha_1$	0	0	0	$\pi_{CG}\alpha_1$	0	0	0	0	0	$\pi_{CA}\alpha_1$	$\pi_{CC}\alpha_2$	*	0	$\pi_{UU}\alpha_2$
UC	0	0	$\pi_{GC}\alpha_1$	$\pi_{UA}\alpha_1$	$\pi_{UG}\alpha_1$	0	0	0	$\pi_{AC}\alpha_1$	0	0	0	$\pi_{CC}\alpha_2$	0	*	$\pi_{UU}\alpha_2$
UU	$\pi_{AU}\alpha_1$	$\pi_{GU}\alpha_1$	0	$\pi_{UA}\alpha_1$	$\pi_{UG}\alpha_1$	0	0	0	0	0	0	0	0	$\pi_{CU}\alpha_2$	$\pi_{UC}\alpha_2$	*

$Q =$

Table C.16: Model 16K

	AU	GU	GC	UA	UG	CG	AA	AG	AC	GA	GG	CA	CC	CU	UC	UU
AU	*	π_{GU}	0	0	0	0	$\pi_{AA}\alpha_1$	$\pi_{AG}\alpha_1$	π_{AC}	0	0	0	0	$\pi_{CU}\alpha_1$	0	$\pi_{UU}\alpha_1$
GU	π_{AU}	*	π_{GC}	0	0	0	0	0	0	$\pi_{GA}\alpha_1$	$\pi_{GG}\alpha_1$	0	0	$\pi_{CU}\alpha_1$	0	$\pi_{UU}\alpha_1$
GC	0	π_{GU}	*	0	0	0	0	0	π_{AC}	$\pi_{GA}\alpha_1$	$\pi_{GG}\alpha_1$	0	$\pi_{CC}\alpha_1$	0	$\pi_{UC}\alpha_1$	0
UA	0	0	0	*	π_{UG}	0	$\pi_{AA}\alpha_1$	0	0	$\pi_{GA}\alpha_1$	0	π_{CA}	0	0	$\pi_{UC}\alpha_1$	$\pi_{UU}\alpha_1$
UG	0	0	0	π_{UA}	*	π_{CG}	0	$\pi_{AG}\alpha_1$	0	0	π_{GG}	0	0	0	$\pi_{UC}\alpha_1$	$\pi_{UU}\alpha_1$
CG	0	0	0	0	π_{UG}	*	0	$\pi_{AG}\alpha_1$	0	0	$\pi_{GG}\alpha_1$	π_{CA}	$\pi_{CC}\alpha_1$	$\pi_{CU}\alpha_1$	0	0
AA	$\pi_{AU}\alpha_1$	0	0	$\pi_{UA}\alpha_1$	0	0	*	π_{AG}	$\pi_{AC}\alpha_1$	π_{GA}	0	$\pi_{CA}\alpha_1$	0	0	0	0
AG	$\pi_{AU}\alpha_1$	0	0	0	$\pi_{UG}\alpha_1$	$\pi_{CG}\alpha_1$	π_{AA}	*	$\pi_{AC}\alpha_1$	0	π_{GG}	0	0	0	0	0
AC	π_{AU}	0	π_{GC}	0	0	0	$\pi_{AA}\alpha_1$	$\pi_{AG}\alpha_1$	*	0	0	0	$\pi_{CC}\alpha_1$	0	$\pi_{UC}\alpha_1$	0
GA	0	$\pi_{GU}\alpha_1$	$\pi_{GC}\alpha_1$	$\pi_{UA}\alpha_1$	0	0	π_{AA}	0	0	*	π_{GG}	$\pi_{CA}\alpha_1$	0	0	0	0
GG	0	$\pi_{GU}\alpha_1$	$\pi_{GC}\alpha_1$	0	π_{UG}	$\pi_{CG}\alpha_1$	0	π_{AG}	0	π_{GA}	*	0	0	0	0	0
CA	0	0	0	π_{UA}	0	π_{CG}	$\pi_{AA}\alpha_1$	0	0	$\pi_{GA}\alpha_1$	0	*	$\pi_{CC}\alpha_1$	$\pi_{CU}\alpha_1$	0	0
CC	0	0	$\pi_{GC}\alpha_1$	0	0	$\pi_{CG}\alpha_1$	0	0	$\pi_{AC}\alpha_1$	0	0	$\pi_{CA}\alpha_1$	*	π_{CU}	π_{UC}	0
CU	$\pi_{AU}\alpha_1$	$\pi_{GU}\alpha_1$	0	0	0	$\pi_{CG}\alpha_1$	0	0	0	0	0	$\pi_{CA}\alpha_1$	π_{CC}	*	0	π_{UU}
UC	0	0	$\pi_{GC}\alpha_1$	$\pi_{UA}\alpha_1$	$\pi_{UG}\alpha_1$	0	0	0	$\pi_{AC}\alpha_1$	0	0	0	π_{CC}	0	*	π_{UU}
UU	$\pi_{AU}\alpha_1$	$\pi_{GU}\alpha_1$	0	$\pi_{UA}\alpha_1$	$\pi_{UG}\alpha_1$	0	0	0	0	0	0	0	0	π_{CU}	π_{UC}	*

$Q =$

Appendix D

Modifications to the PHASE software

Summary of changes from version 2.0

Version 2.1 of PHASE consists of improvements to the stability of the programs when using relatively short alignments and/or highly parametrised models, and some new functionality in the maximum likelihood programs (*mlphase* and *optimizer*). PHASE is no longer actively maintained, and I describe an unofficial version 2.1 of PHASE that has not (yet) been sanctioned by the original authors. The changes are relatively small (though not, I think, unimportant), and I would like to acknowledge the efforts of the original authors of PHASE, and make it clear that credit for the software chiefly rests with them (Jow *et al.*, 2002; Hudelot *et al.*, 2003; Gibson *et al.*, 2005; Telford *et al.*, 2005; Gowri-Shankar and Rattray, 2006, 2007); see also the Acknowledgements section of the PHASE 2.0 manual.

This unofficial version 2.1 of Phase is freely available at <http://www.monkeyshines.co.uk/phase>.

New functionality: Empirical frequencies

The parameter optimization in version 2.0 of PHASE's maximum likelihood (ML) programs (*mlphase* and *optimizer*) estimates (di)nucleotide frequency parameters via ML rather than using empirical frequencies. (Amino acid and codon models have an *Empirical Values* option, for specifying a file containing empirical frequencies, such as the JTT or WAG matrices.) ML estimates are potentially useful in evolutionary models in general, and in RNA models in particular, where the original authors of PHASE found that the empirical and ML frequencies of non-canonical base pairs differed significantly (Jow *et al.*, 2002; Hudelot *et al.*, 2003). Nonetheless, empirical frequencies

are widely used, and can provide sufficiently good estimates of the values to be useful. Having the option to use empirical frequencies allows one to compare the effect on likelihood values of empirical versus ML frequencies, and makes it possible to compare models of evolution with different numbers of states.

Control file settings

The default in PHASE 2.1 is to use ML estimates of frequency parameters; to use frequencies calculated from the alignment, the following line should be added to the **MODEL** block of the control file:

```
Empirical frequencies = yes
```

If a mixed model is used, the setting applies to all of the models, and thus must appear in the **MODEL** block, rather than any of the **MODEL i** blocks (where i is the model number). ML estimation of frequencies can be explicitly indicated by setting the value of *Empirical frequencies* to 'no'. Note that, as with all options in PHASE control files, *Empirical frequencies* is case-sensitive.

Calculation details

The calculation of empirical frequencies is not always as straightforward as it may seem. The mathematics is trivial for ungapped alignments of unambiguous nucleotide sequences. Handling gaps and ambiguities boils down to a single problem, since the standard way to deal with gaps in ML phylogeny software is to treat them as missing data, effectively as an ambiguous nucleotide. (This approach, attributed to Felsenstein (2005), is often acknowledged as less than ideal, but the only practicable alternative is to discard columns with gaps, potentially discarding large amounts of data.) If there is an ambiguous nucleotide in an alignment, an 'R' representing a purine for example, it should contribute to the frequency of both purines, A and G. However, incrementing the counts of each equally, by 0.5, is unsatisfactory, since A and G will rarely have exactly equal frequencies. The approach that is typically adopted is to weight the increment by the frequency of the unambiguous nucleotides.

For example, in the following alignment,

```
Human ACR
Loris AAG
Potto AAG
```

the frequencies are calculated as follows:

$$\begin{aligned}\pi(A) &= \frac{\#A}{9} + \left(\frac{\#A}{\#A+\#G} \frac{\#R}{9}\right) = \frac{5}{9} + \left(\frac{5}{7} \frac{1}{9}\right) = 0.63492 \\ \pi(C) &= \frac{\#C}{9} = \frac{1}{9} = 0.11111 \\ \pi(G) &= \frac{\#G}{9} + \left(\frac{\#G}{\#A+\#G} \frac{\#R}{9}\right) = \frac{2}{9} + \left(\frac{2}{7} \frac{1}{9}\right) = 0.25397 \\ \pi(T) &= 0\end{aligned}$$

When performing these calculations, gaps are ignored. One might think that if we are treating gaps as ambiguities for the purposes of likelihood calculation, then we ought to do so in the frequency calculations too; but that approach is a pragmatic workaround, and it makes more sense to base calculations on just the data that we see. In practice, the difference between including or ignoring gaps is negligible; and is, in fact, non-existent if there are no ‘true’ (i.e. non-gap) ambiguity characters in the alignment, since any weights that arise from ‘ambiguous’ gaps are exactly proportional to the nucleotide frequencies to which they are added.

Dealing with gaps in RNA alignments is slightly more troublesome, since one member of a dinucleotide pair might be a gap, while the other is a nucleotide. (If both members are gaps, then the dinucleotide can be ignored, analogous to the nucleotide case). There are three ways to approach this problem (Table D.1).

Method	Pros	Cons
Treat as “-”, i.e. ignore base pair	Simple	Discards information
Treat as “AN”, i.e. convert gap to ambiguity	Makes use of the knowledge of the non-gap base	Inconsistent with the calculation of single nucleotide frequencies
Treat as two unpaired bases, “A” and “-”	Sensible in biological terms: if a base pair loses its partner, how can it still be treated as part of a pair?	Impracticable to treat a base as paired in one aligned sequence and unpaired in another

Table D.1: Options for calculating frequencies of base pairs in which one member is a gap; the example of “A-” is illustrated. It is assumed that the gaps are generated through biological processes, rather than representing missing data.

In PHASE 2.1, the first approach in Table D.1 is taken: any dinucleotides that contain a gap are ignored when calculating empirical frequencies. Treating the gap as an ambiguity is the only alternative that could be implemented, and this is conceptually problematic because it suggests that it is useful to include data on one half of a pair, and it is unclear whether this is biologically sensible. It could be argued that dinucleotides with gaps should be replaced with a pair of gaps for the likelihood calculations as well as the empirical frequency calculations; PHASE 2.1 will not do this automatically, but this pre-processing step is performed for the analyses described in Chapter 5.

An additional complication arises if a frequency is zero, because a pseudocount approach is used in PHASE 2.1 so that these frequencies have a very small non-zero value (for details, see the section on Zero-valued parameters). The frequencies will be different depending on whether the pseudocount is added before or after ambiguous nucleotides are counted, but in practice the difference is so small it is extremely unlikely to impact the results of the inference. In the PHASE calculations, the pseudocount is added after the ambiguities are handled.

New functionality: Mapping 7-state likelihoods to 16-state space

Until now, there was no easy way to compare the likelihoods between DNA and RNA models with different state spaces. As was shown in Chapter 5, the likelihoods of 4-state nucleotide and 16-state dinucleotide models are, in fact, directly comparable; and a likelihood correction is easily calculated to make 7-state models equivalent to a 16-state model. The *optimizer* program in PHASE 2.1 reports the likelihood correction value in its output if a 7-state model is used, for both the empirical and equal frequency variations.

New functionality: Additional models

DNA model: F81

Earlier versions of PHASE omitted the F81 model (Felsenstein, 1981), which allows unequal equilibrium frequencies, and assumes a single rate of exchangeability. More complex models are often used now, but it has been added for completeness.

RNA model: RNA7G

The RNA7G model takes the process of parameter generalisation to its natural conclusion within the set of 7-state models. It combines the 7E and 7F models, and thus has base pair symmetry in the frequency parameters, and a mismatch rate and a single transition rate; double substitutions are not permitted. With 4 free parameters, it is the simplest available 7-state model.

Bug fixes: Program stability

Compilation

Since PHASE 2.0 was written, new versions of GCC have been released which are stricter about requiring headers to be explicitly included; consequently, PHASE 2.0 fails to compile unless you use an old (3.x) version of GCC. The source files in version 2.1 have been updated appropriately, so that compilation should not be a problem; the latest version of GCC that has been tested is 4.5.3.

Zero-valued parameters

The problem In version 2.0, PHASE instantiated all rate and frequency parameters as zero, and then modified them using empirical values from the sequence alignment; if there were no instances of a particular change in the alignment, the rate would remain at zero. However, PHASE always uses a specific rate as a reference (e.g. $A \leftrightarrow G$ for DNA models), and the program would either hang or crash if that rate was zero, since the normalization process would then be trying to divide by zero. In a similar manner, if a character is not seen in the alignment, its frequency would be zero, and PHASE would crash or hang when calculating rate heterogeneity with the gamma distribution. One might ask how likely it is that a rate or frequency never occurs in an alignment: it is not common (which is why most users will not have encountered this problem), but it is possible, particularly for relatively short alignments and highly-parametrised models, such as the 16-state RNA models.

Even if the zero-valued parameters did not cause the program to fail, the ML optimization code is written such that these parameters could never escape from a value of zero. Implicitly fixing one or more parameters (at zero or any other value) is undesirable, because it limits the likelihood values that can be reached by the optimization.

There is also a more abstract reason for disallowing zero values, which is that in creating mathematical models of evolution, we should allow the possibility of rare events, albeit with extremely small probabilities.

The solution For the pragmatic and theoretical reasons outlined above, model parameters cannot be zero-valued in version 2.1 of PHASE. In the case of rates, this is achieved by instantiating the parameters to be $1e-6$, a value that is non-negligible to the optimization routine, but only just, given the accuracy with which PHASE performs numerical calculations. For frequency parameters, a small value (again, $1e-6$) is added to each frequency, and then all frequencies are adjusted so that they sum to 1. There are more sophisticated ways to perform pseudocount calculations such as this, but since it just represents the starting point for optimization, a more complicated process is unnecessary. (And if empirical frequencies are being used in the analysis, the value is, for practical purposes, essentially zero.) The amino acid and codon models in PHASE 2.0 already have a pseudocount added to the frequency parameters, so the changes mentioned above were not implemented for these models (non-zero rate parameters *were* implemented, though).

Postscript

In fact data itself was soulful and glowing, a dynamic aspect of the life process. This was the eloquence of alphabets and numeric systems, now fully realized in electronic form, in the zero-oneness of the world, the digital imperative that defined every breath of the planet's living billions. Here was the heave of the biosphere. Our bodies and oceans were here, knowable and whole.

Don DeLillo, *Cosmopolis* (2003)