Codon Usage and Molecular Phylogenetics

Studies of Codon Usage and Molecular Phylogenetics Using Mitochondrial Genomes

By

Wenli Jia, B.Sc., M.Eng.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

©Copyright by Wenli Jia, December 2007

MASTER OF SCIENCE (2007) McMaster University

(Computational Engineering and Science) Hamilton, Ontario

TITLE: Studies of Codon Usage and Molecular Phylogenetics Using Mitochondrial Genomes

AUTHOR: Wenli Jia, B.Sc., M.Eng.

SUPERVISOR: Dr. Paul G. Higgs

NUMBER OF PAGES: xxiv, 170

Abstract

Three pieces of work are contained in this thesis. OGRe is a relational database that stores mitochondrial genomes of animals. The database has been operational for approximately five years and the number of genomes in the database has expanded to over 1000 in this period. However, sometimes, new genomes can not be added to the database because of small errors in the source files. Several improvements to the update method and the organizational structure of OGRe have been done, which are presented in the first part of this thesis.

The second part of this thesis is a study on codon usage in mitochondrial genomes of mammals and fish. Codon usage bias can be caused by mutation and translational selection. In this study, we use some statistical tests and likelihood-based tests to determine which factors are most important in causing codon bias in mitochondrial genomes of mammals and fish. It is found that codon usage patterns seem to be determined principally by complex context-dependent mutational effects.

The third part of this thesis is a phylogenetic study of 159 avian species obtained using mitochondrial rRNA sequences that were provided by Dr. van Tuinen. In this study, two methods are used: one considers sites of sequences as independently evolving; the other includes the secondary structure of rRNAs. Unfortunately, the amount of information in the rRNA sequences seems to be insufficient to determine the whole phylogeny of birds. However, our results make it clear that several traditionally defined orders are polyphyletic and therefore need to be redefined.

Acknowledgements

I would like to thank my supervisor Dr. Paul Higgs for all his invaluable help and directions. I would like to thank Dr. Brain Golding and Dr. James Wadsley for their suggestions about my thesis. I would like to thank Dr. van Tuinen for his provision of avian sequences.

Thanks to Mark Wu for his all help with LATEX and others. More thanks go to Wenqi Ran, Xiaoguang Yang and Wei Xu.

I would also like to take this opportunity to thank my family for their support.

dedicated to my family

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Biology of Mitochondria

Mitochondria are membrane-enclosed organelles found in most eukaryotic cells (Jameson, 2004). Mitochondria are known as "cellular power plant" because they provide energy for all sorts of machinery in cells such as metabolic reactions, substance transports and mechanical work. They convert energy from oxygen and nutrients into adenosine-triphosphate (ATP), and then ATP transports and delivers energy to any activities that consume energy. The number of mitochondria in a cell is related with the function of this cell. For example, muscle cells have more mitochondria than other cells because they need more energy.

Usually, mitochondria are elongated and rod-shaped organelles (Alberts *et al*, 2002). They have double specialized membranes, which have quite different structures and functions (Alberts *et al*, 2002). The outer membrane, which is fairly smooth, encloses the organelle as its shell. It filters out large molecules and allows most small molecules and ions to pass through. The inner membrane, which is continuous and thin surface, contains many functional membrane proteins, critical components and essential enzymes for aerobic respiration and ATP production. The inner membrane is highly convoluted and forms numerous inward folds that are called cristae and greatly increase the total surface area of inner membrane. Like the outer membrane, the inner membrane is especially impermeable to some ions, but it is selectively per-

meable for smaller molecules that are transported by various proteins. The outer and inner membranes form two distinct compartments within mitochondria. The space enclosed by inner membrane is named matrix, which includes genomes, ribosomes, tRNAs and enzymes. The narrow region between the two membranes is called intermembrane space, which also contains enzymes. The structure of mitochondria is shown in Figure1.1. In most cells, mitochondria replication is independent of the cell cycle (Bogenhagen and Clayton, 1977). Mainly based on the energy needs of cells, mitochondria can reproduce themselves semi-autonomously within cells. Like bacterial cells, each mitochondrion doubles in mass then splits into two daughter mitochondria. However, unlike bacterial cells, mitochondria can also fuse with one another (Chan, 2006; Wiesner *et al*, 1992; Alberts *et al*, 2002).

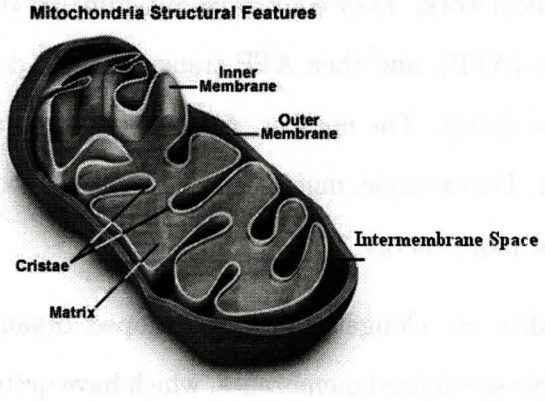

Figure 1.1. The structure of mitochondria (courtesy of Molecular Expressions Website)

Mitochondria contain their own genomes although most DNA in a eukaryotic cell is kept in its nucleus. In most species, mitochondrial genomes are circular and distinct with those in the nucleus. They usually contain many fewer genes than nuclear genomes. Typically, each mitochondrial genome of animals is tightly packed with 13 protein genes, 22 transfer tRNAs, one small ribosomal RNA and one large

ribosomal RNA (Jansen, 2000; Wolstenholme, 1992) Figure1.2 is the diagram of the human mitochondrial genome drawn by OGRe, which is represented as linear rather than circular to save space. Each gene is drawn as a block and the word of the block is the name of the gene. Different types of genes are labeled by different colors: green for proteins, red for tRNAs and blue for rRNAs.

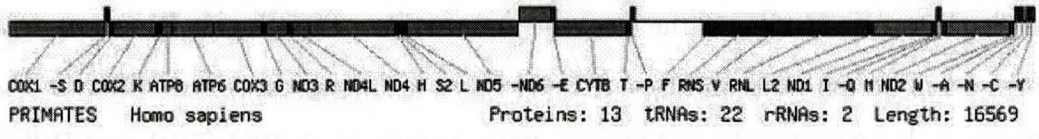

Figure 1.2: The human mitochondrial genome

Although the mitochondrial genome is entirely separate from nuclear DNA, mitochondrial DNA (mtDNA) and nuclear DNA are functionally interdependent (Castro, 1998) Both of them encode mitochondrial proteins although the contribution proteins coded by nuclear DNA is much larger than that coded by mtDNA (Fernández-Silva, 2003). Like nuclear DNA, mtDNA has two strands. However, the two strands have uneven density (Falkenberg *et al*, 2007), so they are denoted as heavy (H) strand and light (L) strand. Both strands have different origins of replication and their replications are not synchronously. Hence, the replication of mtDNA is asymmetric. Figure1.3 shows the process. The replication starts at a site called O_H on the H-strand, which is also called the leading strand, and it proceeds unidirectionally (Brown TA *et al*, 2005) New hydrogen bonds are formed between the daughter H-strand and the parental L-strand. As a result, the parental H-strand is displaced and becomes a single strand. When the formation of the daughter H-strand reaches the origin of the L-strand, O_L, which is about two-thirds of the away around the circular genome, the replication of L-strand, which is termed the lagging strand, is initiated. From this point, the L-strand is replicated on the single-stranded template

in the opposite direction. Hence, when one-third of the lagging strand is replicated, the replication of the leading strand is finished and the daughter H-strand and the parental L-strand separate from the replication of the lagging strand. Then, a daughter mitochondrial genome is formed, while the daughter molecule with the parental H-strand is delayed (Shadel and Clayton 1997) Due to the asymmetric replication mechanism of mitochondrial genomes, different genes on the parental H-strand have varying amounts of time in the single-stranded state. This time is known as D_{ssH} Different mutation rates and base frequencies arise among genes with different D_{ssH} (Reyes *et al*, 1998; Tanaka and Ozawa, 1994; Jermiin *et al*, 1995; Faith and Pollock, 2003; Urbina *et al*, 2006).

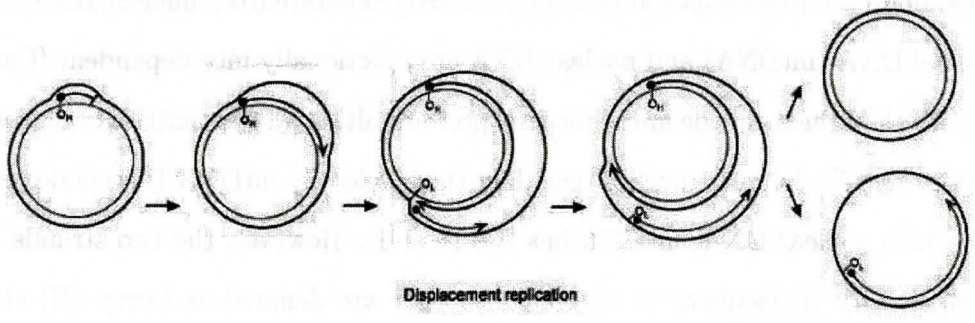

Figure 1.3: The replication of mtDNA (Brown TA *et al*, 2005)

Usually, mitochondria are inherited through the maternal. Hence, they can help researchers to discover the history of female lineage. The famous Mitochondrial Eve is a typical example of this type of analysis. Unlike nuclear DNA, mitochondrial genomes are not highly conserved and have high rates of mutation (Fernández-Silva, 2003). Moreover, mitochondria are nearly lack of the genetic recombination, which is the process of regrouping the maternal and paternal genes. Hence, they become a very useful source for the evolutionary biology studies.

Codon	Aa	Codon	Aa	Codon	Aa	Codon	Aa
UUU	F	UCU	S	UAU	Y	UGU	C
UUC	F	UCC	S	UAC	Y	UGC	C
UUA	L	UCA	S	UAA	*	**UGA**	**W**
UUG	L	UCG	S	UAG	*	UGG	W
CUU	L	CCU	P	CAU	H	CGU	R
CUC	L	CCC	P	CAC	H	CGC	R
CUA	L	CCA	P	CAA	Q	CGA	R
CUG	L	CCG	P	CAG	Q	CGG	R
AUU	I	ACU	T	AAU	N	AGU	S
AUC	I	ACC	T	AAC	N	AGC	S
AUA	**M**	ACA	T	AAA	K	**AGA**	*
AUG	M	ACG	T	AAG	K	**AGG**	*
GUU	V	GCU	A	GAU	D	GGU	G
GUC	V	GCC	A	GAC	D	GGC	G
GUA	V	GCA	A	GAA	E	GGA	G
GUG	V	GCG	A	GAG	E	GGG	G

Table 1.1: The genetic code of vertebrate mitochondria

1.2 Codon usage

Codons are the sequences of three adjacent nucleotides in a DNA or messenger RNA molecule, each of which codes for a specific amino acid in a protein. Hence, they work as a manual for the construction of proteins. The translational map between codons and amino acids is called the genetic code. Most species use the same genetic code that is known as the canonical code. However, it was found that mitochondrial genomes use a different genetic code (Osawa, 1995; Knight *et al*, 2001; Sengupta *et al*, 2007). Furthermore, mitochondria from different groups of organisms sometimes use different codes. The code that is most relevant to the work in this thesis is the vertebrate mitochondrial code. This mitochondrial code is shown in Table1.1, in which the bold codons have been reassigned relative to the canonical code. In the canonical code, UGA is a stop codon (*), AUA codes for Ile (I), AGA and AGG code for Arg (R).

There are 64 possible codons but only 20 amino acids. Thus, one amino acid can be encoded by two or more codons that are known as synonymous codons. Typically, degenerate codons are different in their third positions, for example, both GAA and GAG specify the amino acid Glu, and all four codons GGU, GGC, GGA and GGG code for Gly. If an amino acid can be specified by a codon with any one of the four nucleotides at its third position, the third position is called four-fold degenerate. Similarly, if a codon with two of the four nucleotides at its third position can specify the same amino acid, the third position is known as two-fold degenerate.

After the discovery of the redundancy of the genetic code and before the 1980s, it was often believed that synonymous codons were used randomly in a genome. Sometimes, it is also thought that codon usage, which is the frequency of each codon in a gene or genome, was uniform. However, more and more evidence shows that synonymous codons are used with unequal frequencies and codon usage is non-random in most organisms. If some codons are particularly preferred over others that encode the same amino acid, this is called codon usage bias.

Codon usage bias might reflect a balance between mutation and translation selection. One usual type of selection is translational efficiency, *i.e.* an organism prefers to use codons that are more rapidly translated in order to reduce the time and effort spent on translation. It has been shown in several species, *e.g. E. coli, Drosophila* and *Caenorhabditis elegans* (Sharp *et al*, 1988; Akashi, 2003; dos Reis *et al*, 2003) and is most important for organisms with rapid growth rate (Sharp *et al*, 2005) because the time saving is more significant for such species. As well, within any genome, translational efficiency is more important in highly expressed genes because a large fraction of the translational effort of the cell can be spent on making large numbers of copies of a relatively small number of proteins. Another type of selection is translational accuracy, which is that codons are preferred because they have lower probabilities of

mistranslation (Akashi, 1994; Stoletzki and Eyre-Walker, 2007). On the other hand, one of the mutation forces is context dependent mutation, which means the rate of mutation from any one base to any other is influenced by bases at the neighbouring sites. This kind of mutation has been detected in many different genomes (Karlin and Mrazek, 1997; Shioiri and Takahata, 2001), and studied specifically in humans (Karlin and Mrazek, 1996), Drosophila (Antezana and Kreitman, 1999; Fedorov *et al*, 2002), and Arabidopsis (Morton and Wright, 2007).

1.3 Relational Databases

Relational databases were invented in 1970 when a researcher (Codd, 1970) at IBM proposed the idea. Relational database are databases that stores data in a set of logical relations, which are represented as logically related tables. They allow their data to be extracted and recombined in many different ways without needing to change the existing tables. Besides, relational databases can be extended easily. New relations can be added into the original databases without modifying the old relations and applications.

In a relational database, each relation, *i.e.* each table, is composed of one or more fields, which are described as columns. The fields are the basic units of data storage and keep attributes of each entity. Each record, which is represented as a row, contains one value for each field. Values in all these fields together provide a unique description of each particular record. Besides, each relation always has a primary key, which should be unique because it is used to identify each record in the table (Connolly and Begg, 1998). The primary key may consist of a single field or a combination of multiple fields. Table1.2 is an example of a relation, whose primary key is the combination of the GENOME_CODE, CODON and STRAND fields. Moreover,

a relation can contain a foreign key, which is a referential link between two tables. A foreign key can be established by adding one column or columns in one table, which point to the primary key in the other table. The former table is denoted as the child table and the latter table is named the parent table. One relation may possess several foreign keys that can be related to different tables. The purpose of foreign keys is to ensure that data in the two related tables are consistent. For example, there are two tables: a STUDENT table that contains all data of students and a CLASS table that contains all classes that our school provide. In this case, let us suppose students are only permitted to choose classes listed in the CLASS table. To satisfy this requirement, a foreign key can be created on the STUDENT table and make it correlate to the primary key of the CLASS table. Hence, the foreign key makes sure that classes that do not appear in the CLASS table are not permitted to be saved in the STUDENT table. Furthermore, foreign keys can be configured to do a cascade deletion. This means that when an entity in the parent table is deleted, its matching entities in the child table will automatically be deleted (Nijssen and Halpin, 1989). Meanwhile, foreign keys also put a constraint on the primary key in the parent table. In other words, a value of the primary key can not be deleted or modified until it does not have any corresponding records in the child table.

To support any relational database, there must be a system to manage its data, which is called a database management system (DBMS). A DBMS usually provides various commands so that users can quickly query, reassemble and retrieve of data in several tables at the same time. Besides, tables in a relational database can be updated, modified and deleted using commands. At present, relational databases have become one of the most common types of databases in use, which have covered thousands of topics like medicine, law, business, news, etc. Our database, OGRe, is

genome_code	codon	strand	usage
URASP1MIT	TAG	+	1
URASP1MIT	GCC	+	87
URASP1MIT	TAG	-	0
URASP1MIT	GCC	-	0
HOMSAPMIT	TAG	+	3
HOMSAPMIT	GCC	+	123
HOMSAPMIT	TAG	-	0
HOMSAPMIT	GCC	-	0

Table 1.2: an example of table

one of them and can be accessed publicly, which contains a great deal of biological data.

1.4 The OGRe Database

OGRe is designed to organize and study the increasing mitochondrial genomes, which provides simple tools to access the sequences and their related information for comparative analyses. The basis for the information in OGRe is GenBank (http://www.ncbi.nlm.nih.gov/Genbank/), which contains a list of mitochondrial genomes and is one of the three most popular primary databases. The three databases exchange their data daily to ensure that each of them contains the comprehensive information of all public sequence (Benson *et al*, 1994). The information of genomes in GenBank can only be saved as flat files that are very difficult to be retrieved for some specific information and maintained for some common mistakes. For example, if the name of a taxonomic organism is wrong, all files that contain this organism have to be found, and then the name need to be changed in every file. On the contrary, if all these

information is stored in a relational database and the name is stored in an individual table, the work is just to change one record in that table. Obviously, relational databases are much easier to maintain than flat files, which is one of reasons why OGRe is a relational database.

OGRe stands for Organellar Genome Retrieval. The database was designed originally by Daniel Jameson in 2001 at University of Manchester and the version at McMaster was set up by Bin Tang in 2003. At the sequence level, any single or multiple mitochondrial genes of selected organisms can be obtained and downloaded on the web site of OGRe (http://ogre.mcmaster.ca). At the level of separate nucleotides, the database can show the information on base frequencies and codon usage frequencies of any group of species (Jameson *et al*, 2003). Several phylogenetic studies on mammals (Jow *et al*, 2002; Hudelot *et al*, 2003; Gibson *et al*, 2004) and on arthropods (Xu *et al*, 2006) have been done with OGRe. As well, it can display statistics of bases, codon usages and amino acids for each strand of any selected set of species. This is interesting because Urbina *et al* (2006) found that the frequencies of amino acids are related to their physical properties. At the level of complete genomes, OGRe provides several visualization tools. The database can draw diagrams that show gene order of any selected species and illustrate their genome arrangement automatically. Similarly, it can generate figures of gene order for any two comparable genomes, which highlight the sections of break points in the gene order by using colours (Jameson *et al*, 2003). Figure1.4 is an example of a pairwise genome comparison. This tool has been used to study the gene rearrangements in arthropods, whose rate is found to be related to the rate of sequence evolution (Higgs *et al*, 2003; Xu *et al*, 2006).

At present, OGRe focuses on metazoan species, *i.e.* animal species, and only includes mitochondrial genomes that have complete sequences (Jameson *et al*, 2003). At the time of original publication there were approximately 250 species, but this

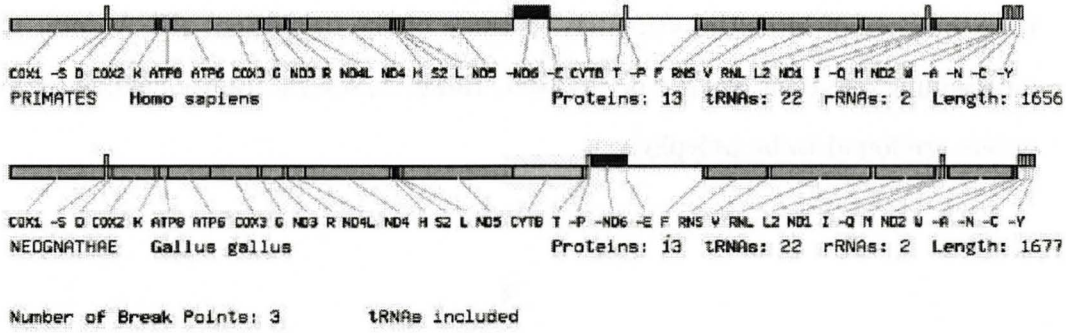

Figure 1.4: An example of a pairwise genome comparison produced by OGRe where tRNA genes are included

has increased to more than 1000 species (as of December 2007) due to the regular updates that I have been making during the period of my thesis. Species are classified by a taxonomic system so that the selection of close species is easy. The open source and powerful PostgreSQL is chosen as the database management system of OGRe because it supports the standard of structured query language (SQL), provides a well structured Perl interface and is able to generate HTML tables (Jameson, 2004)

1.5 Aims of This Thesis

In order to make OGRe easier to be managed and added new genomes, several improvements have been done to its structure and the method of update. In Chapter2, I will present these problems and their improvements. In Chapter3, a study of codon usage in mitochondrial genomes will be presented, which only focus on a part of mammalian and fish species stored in OGRe. This chapter will contain the statistic tests and likelihood-based tests, which we used to examine the presence of mutation and translational selection, and the discussion about what we obtained. In Chapter4, I will present a phylogenetic study on rRNA genes of 159 avian species, which is

a study of the evolution of related organisms. This study tried two methods on the sequences of birds.Unfortunately, because of the insufficient rRNA data, many outstanding questions can not be resolved. However, several traditional defined avian orders are found to be polyphyly.

Chapter 2

The Improvements of OGRe

This chapter deals with improvements to the organizational structure of OGRe and to the method by which new sequences are added to the database. These are intended to make OGRe easier to manage and keep up to date with the large numbers of new genomes that are being sequenced.

2.1 Source Files of OGRe

The data of OGRe is obtained from flat files of complete mitochondrial genomes, which are downloaded from GenBank. The flat files are specific structured and can be easily translated manually. Each flat file consists of a sequence and a brief description of an organism. Figure2.1 is an example of a flat file. In each flat file, the left tags of fields indicate what categories the particular information on the right side belongs to. The length of sequence is listed in the first line of each flat file. In the ACCESSION field, the eight-character alphanumeric identifier is the accession number, which is issued uniquely for every record in GenBank and can not be modified (Benson DA, 1994). Usually, an accession number is a combination of letters and numbers and its length depends on the type of its sequence record. For animals, their accession numbers are composed of two uppercase letters followed by an underscore and six digits, for example, NC_001056. The SOURCE field usually includes the common name of the source organism. The ORGANISM field contains the scientific name of the source organism and its taxonomic lineage. The REFERENCE field lists the

related publications about the source sequence, which is written by the submitters of the source sequence. The FEATURES field is a table of protein, tRNA and rRNA genes and other regions such as D_loops, source and control region. The table provides the name, type and region of each feature. Meanwhile, the table shows that which strand each feature is on. At the end of a flat file, the sequence data is listed in the ORIGIN field.

2.2 The Schema of the original OGRe

Flat files of GenBank show that there are several common types of data for each genome. Each file possesses the basic information of a genome, such as its length, its accession number, its brief description, its type, its sequence, its source organism and the taxonomic lineage of its source organism. Besides, the GenBank file contains the important data of protein, tRNA and rRNA genes, etc. Because the basic information of each genome is only related to the genome, the original OGRe uses one table that is called GENOME to store them. On the other hand, because one source species may potentially have two or more genomes, the information of source organisms is kept in another table that is named SPECIES and contains the scientific name, common name and taxonomic classification of source species. Each genome has several features and each feature also has several types of basic information, so the basic data of features, such as their names and types, are stored in a table, which is called FEATURE. Both the SPECIES table and the FEATURE table are related to The GENOME table. The SPECIES table is linked to the GENOME table by their common field, SPECIES_CODE, and the FEATURE table is linked by the same field, GENOME_CODE. Indeed, the two fields perform as foreign keys. The three tables constitute the core of the original OGRe.

```
LOCUS       NC_006131              16663 bp    DNA     circular VRT 12-JAN-200
DEFINITION  Acanthogobius hasta mitochondrion, complete genome.
ACCESSION   NC_006131
VERSION     NC_006131.1  GI:51101178
KEYWORDS    .
SOURCE      mitochondrion Acanthogobius hasta (javeline goby)
  ORGANISM  Acanthogobius hasta
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleostei;
            Acanthomorpha; Acanthopterygii; Percomorpha; Perciformes;
            Gobioidei; Gobiidae; Acanthogobius.
REFERENCE   1  (bases 1 to 16663)
  AUTHORS   Kim,I.C., Kweon,H.S., Kim,Y.J., Kim,C.B., Gye,M.C., Lee,W.O.,
            Lee,Y.S. and Lee,J.S.
  TITLE     The complete mitochondrial genome of the javeline goby
            Acanthogobius hasta (Perciformes, Gobiidae) and phylogenetic
            considerations
  JOURNAL   Gene 336 (2), 147-153 (2004)
   PUBMED   15246526
REFERENCE   2  (bases 1 to 16663)
  AUTHORS   .
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (21-SEP-2004) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   3  (bases 1 to 16663)
  AUTHORS   Kim,I.-C., Kim,C.-B., Kim,Y.J. and Lee,J.-S.
  TITLE     Direct Submission
  JOURNAL   Submitted (24-NOV-2003) Dept. of Environmental Science, Hanyang
            University, Graduate School, Seoul 133-791, South Korea
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from AY486321.
FEATURES             Location/Qualifiers
     source          1..16663
                     /organism="Acanthogobius hasta"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:267130"
                     /common="javeline goby"
     tRNA            1..68
                     /product="tRNA-Phe"
     rRNA            69..1011
                     /product="s-rRNA"
     tRNA            1013..1084
                     /product="tRNA-Val"
     rRNA            1086..2769
                     /product="l-rRNA"
     tRNA            2770..2843
                     /product="tRNA-Leu"
                     /note="codons recognized: UUR"
     gene            2844..3818
                     /gene="ND1"
                     /db_xref="GeneID:2943469"
     . . . . . .
ORIGIN
        1 gccaacgtag cttaattaaa gcataacact gaagatgtta agatggaccc tagaaagtct
       61 cgttagcaca aaagcttggt cctgactttt ctgtcagctt tggttagact tatacatgca
      121 agtatccgca cccctgtgag aatgccctac acactcccaa acggagtaaa ggagcaggta
      181 ttaggcacga ccacaagtca gcccatgacg ccttgtttag ccacacccta aagggaactc
      241 agcagtaata aacattaagc aacaagtgca aacttgactt aattaagacc aattagggcc
      301 ggtaaaactc gtgccagcca ccgcggttat acgagggggcc caagttgacg gacaccggca
      361 taaaatgtgg taagtactaa aatatactaa agccgaacac cttcaagact gttataagtt
      . . . . . .
//
```

Figure 2.1: An example of a GenBank flat file

In addition, the original OGRe also contains several other tables. The CLASSIFICATION table is used to store the taxonomic lineage of source organisms, which is designed as a hierarchical two-dimensional data structure. This table has an attribute called GROUP_NAME, which is pointed by the SPECIES table so that the lineages and species are linked together. When a feature is a tRNA, the information of the tRNA is saved in a table named TRNA, which contains amino acids, anticodons and codons. Because a single feature may have multiple regions that can be jointed together later, the locations of features have to be stored in different table. The FEATURE_LOCATION table is designed to satisfy this requirement, which contains the start and stop positions of features. Both the TRNA table and FEATURE_LOCATION table are linked to the FEATURE table by the FEATURE_ID field. Besides, the TRANSLATION table stores the translation code used for particular genomes. The CITATIONS table keeps the MEDLINE codes of Citations. The FEATURE_DESCRIPTION, TRANSDESCS and AACIDS tables contain the general descriptions of features, genetic code and amino acids that would be commonly found within the database.

Besides the data that extracted from the flat files of GenBank, OGRe includes other information created by ourselves as well, which is based on the application and performance of OGRe. The GENE_ORDER and GENE_ORDER_NOTRNA tables are designed to store each gene order, a representative genome of each gene order and the number of genomes that each gene order has. The CODON_USAGE table is used to save codons, codon usages and strands of each genome. For the sequence in each flat file, it is more reasonable to save it in files outside OGRe than within the database. Hence, the FILEINDEX table holds the name of the genomic data file and a numerical offset to indicate where the relevant sequence starts in that file, which is directly related to the GENOME table. Figure2.2 is the schema of the original

OGRe, which is a structural description of tables and the relationships among them in the database.

In order to update OGRe, a few specific programs are applied to read the flat files of GenBank, which are called parsers by us and written in Perl because Perl is a very powerfully text processing language (Larry *et al*, 2000). According to the flat files, the parsers generate a separate tab-delimited text file for each table and then load them into the corresponding tables of OGRe.

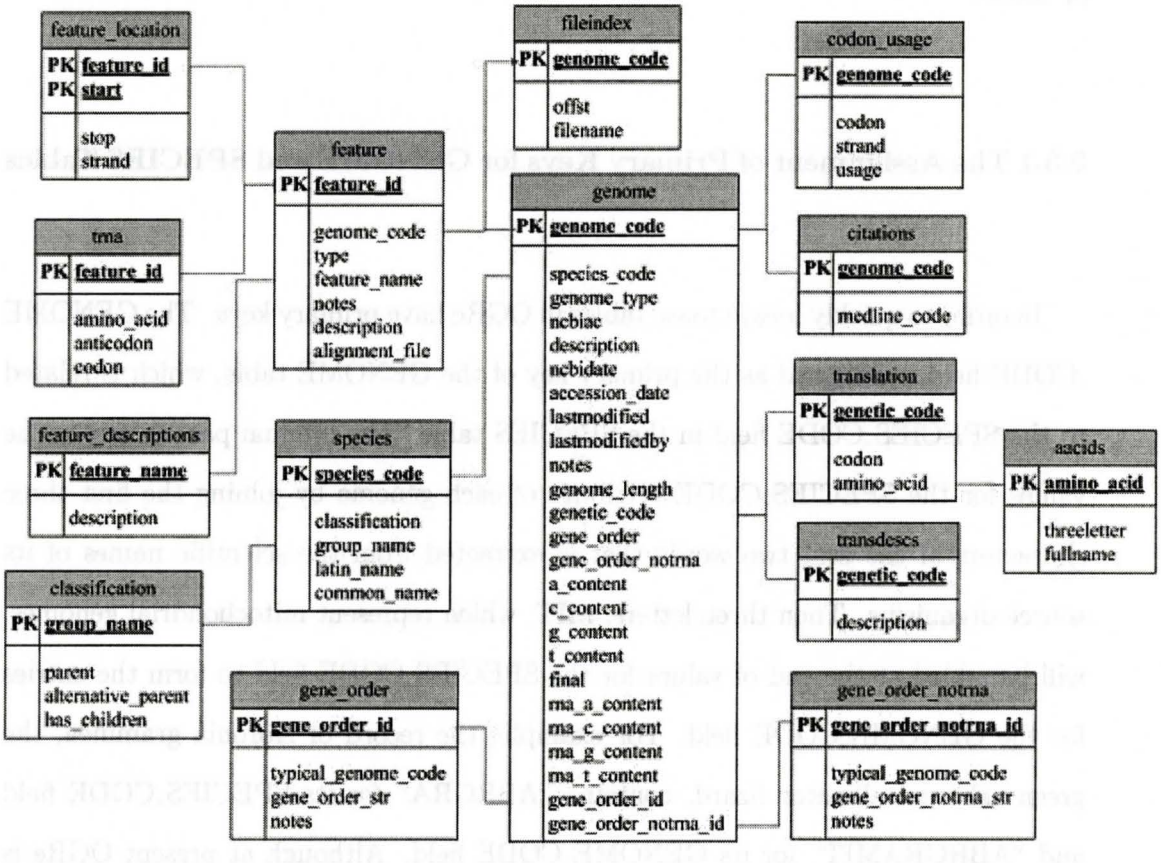

Figure 2.2: The schema of the original OGRe

2.3 The Improvements of OGRe

In principle, the flat files of GenBank are well structured. However, many exceptions exist in the files, such as alternative names for features like trns for tRNA-Ser (AGY), trnl for tRNA-Leu (UUR) and COI for COX1. Each time, the abnormities cause several errors to happen during the update of OGRe. Although the original parsers can deal with some of the exceptions, there are still several other exceptions that can not be solved properly. Therefore, the loading procedure of OGRe needs to be redesigned and the parsers for the update process need to be improved. Meanwhile, the structure of OGRe needs to be simplified by eliminating unnecessary attributes or tables.

2.3.1 The Assignment of Primary Keys for GENOME and SPECIES Tables

In order to quickly locate rows, tables in OGRe have primary keys. The GENOME_CODE field is designed as the primary key of the GENOME table, which is related to the SPECIES_CODE field in the SPECIES table. The original parsers create the values for the SPECIES_CODE attribute of each genome by joining the first three characters of the first two words that is extracted from the scientific names of its source organisms. Then three letters, MIT, which represent mitochondrial genomes, will be added at the end of values for the SPECIES_CODE field to form the entries for the GENOME_CODE field. For example, the record of *Abronia graminea*, the green arboreal alligator lizard, contains "ABRGRA" for its SPECIES_CODE field and "ABRGRAMIT" for its GENOME_CODE field. Although at present OGRe is limited to mitochondrial genomes, it is possible that OGRe will be expanded to other organelle genomes. In that case, one species may possess two or more genomes. Hence,

the GENOME_CODE and SPECIES_CODE attributes are designed to distinguish species and genomes.

However, for the entries in the SPECIES_CODE and GENOME_CODE fields, the original parsers don't consider the case that the length of words is less than three and don't define what the values should be if this kind of exceptions occurs. The improved parsers include three steps to fix this type of problems. First, the length of each word will be checked and then if only one or two characters can be extracted from the flat files, each empty position will be filled by '_' to form temporary entries. Then the values will be modified to what they should be by hand. For example, millipede, whose scientific name is *Thyropygus sp. DVL-2001*, has 'sp' as the second word in its scientific name. When this genome was inserted into OGRe, its temporary entries for the two fields were 'THYSP_' and 'THYSP_MIT' in the individual tab-delimited text files. Then they were changed to be THYSP1' and 'THYSP1MIT' after all similar values contained in OGRe were checked in order to make sure that the new values were unique in OGRe.

Because more and more genomes with similar scientific names are kept in GenBank and the parsers only take the first three characters from two words for the SPECIES_CODE and GENOME_CODE fields, it becomes very difficult to keep the entries for the two fields unique, which is required by the rule of primary keys. Actually, before the parsers are improved, the update of OGRe has failed many times because of two identical entries exist. To solve this type of problems, first, the new parsers will generate values for the two fields according to the basic rule, *i.e.* the joining of three characters. Then the parsers will compare the new entries with those already in OGRe. If the same entities are found in OGRe, the parsers will print the basic information of the new genomes and those in OGRe into a text file. Then the file will be checked and new unique entries for the new genomes can be determined. For

example, when cochin-Chinese red jungle fowl, whose scientific name is *Gallus gallus gallus*, is added to OGRe, temporary entities, 'GALGAL' and 'GALGALMIT' were created. Meanwhile, the related information of this species and chicken, whose scientific name is *Gallus gallus*, were printed into the text file. After the existing entries in OGRe were checked, new values, 'GALGA2' and 'GALGA2MIT', were assigned for the cochin-Chinese red jungle fowl.

2.3.2 The Update of The Classification Table

Although the taxonomic lineages of genomes have been saved as a single attribute, CLASSIFICATION, in the SPECIES table, they are not dynamic. Hence, the CLASSIFICATION table is included in OGRe and saves concise lineages, which is in a hierarchical structure, in other words, a treelike structure. Except the leaf tips, each level of records in such a structure can branch off into records in the lower level. Usually, representative biological taxa are chosen to be saved in the CLASSIFICATION table. For example, the classification of human is Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo. The brief lineage in the CLASSIFICATION table is Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Primates. In the CLASSIFICATION table, each record stands for a taxon that is stored in the GROUP_NAME attribute. At present, because OGRe is limited to animals, the taxon, METAZOA, is the root of all lineages in this table. So except the root, each record has a parent taxon that is actually the previous one of the taxon represented by this record. Meanwhile, except the leaf tips, each record has one or more child taxa, which are the following ones. All the parent and child taxa are kept in the CLASSIFICATION table. A part of the CLASSIFICATION table is shown in Table 2.1. The PARENT attribute is used to store the parent taxa and the HAS_CHILDREN

group_name	parent	has_children
METAZOA		True
CHORDATA	METAZOA	True
VERTEBRATA	CHORDATA	True
MAMMALIA	VERTEBRATA	True
METATHERIA	MAMMALIA	False

Table 2.1: A part of the CLASSIFICATION table

field is used to flag if records have child taxa. If a record has child taxa, the value for HAS_CHILDREN of this record will be set as TRUE, otherwise, it will be set as FALSE. The records with FALSE in the HAS_CHILDREN field are called leaf tips, which is the link between the CLASSIFICATION and SPECIES table and each species should belong to one leaf tip taxon. Figure2.3 shows the tree-structured taxa and the number of organisms within each leaf tip taxon, where the taxa listed in Table2.1 is in red.

Although the GROUP attribute in the SPECIES table links the SPECIES and CLASSIFICATION tables, the original parsers don't update the CLASSIFICATION table before they reload the SPECIES table. In fact, every time, the changes of the CLASSIFICATION table are manually checked and reloaded after all other tables are updated. Because, sometimes, there are hundreds of new genomes, it is very hard to find new lineages without mistakes. Therefore, it is very sensible to update the CLASSIFICATION table automatically because the mistakes can be reduced dramatically. The new parsers are designed to extract every taxon from the classifications of new genomes in the flat files of GenBank and then compare these taxa with those already in OGRe. The comparison of the lineage of each genome starts from its first taxon and follows its order. If the last taxon that can be found in OGRe is a leaf tip taxon, then the comparison of another lineage starts. If the last taxon that can be found in OGRe is not a leaf tip taxon, the new parsers will treat its following

taxon along the lineage as a new taxon and print its default name and related information into text files. After all lineages are examined and the final records of new taxa are ready, then the CLASSIFICATION table will be updated. For example, the classification of the thorny-headed worm, *Leptorhynchoides thecatus*, is Eukaryota; Metazoa; Acanthocephala; Palaeacanthocephala; Echinorhynchida; hadinorhynchidae; Leptorhynchoides. Before it was added into OGRe, the last taxon that could be found in OGRe was Metazoa. Since Metazoa is not a leaf tip taxon, a new taxon, Acanthocephala, had to be added into the CLASSIFICATION table, which is set as a leaf tip taxon. The species, *Leptorhynchoides thecatus*, is jointed with the taxon, Acanthocephala.

2.3.3 The Improvement of Sequence Files and Database Structure

As mentioned above, the sequences are saved in a large file and the FILEINDEX table in OGRe is used to holds the start position of each genome in the large file. During the update of OGRe, the file also needs to be appended the sequences of new genomes. However, its size is over 10.0 M and still keeps increasing. So it takes a long time to open and read the file. Moreover, it is very difficult to find the sequences of some genomes. Therefore, the sequences of genomes are redesigned to be saved in individual files, one file for each genome. The improved parsers define the names of the separate files as their values of GENOME_CODE followed by 'fasta' as the extended name. All sequence files of new genomes are generated automatically by the new parsers and kept in a directory that is called SEQUENCE so that they can be divided from non-data files and managed easily. Thus, the FILEINDEX table is removed from OGRe.

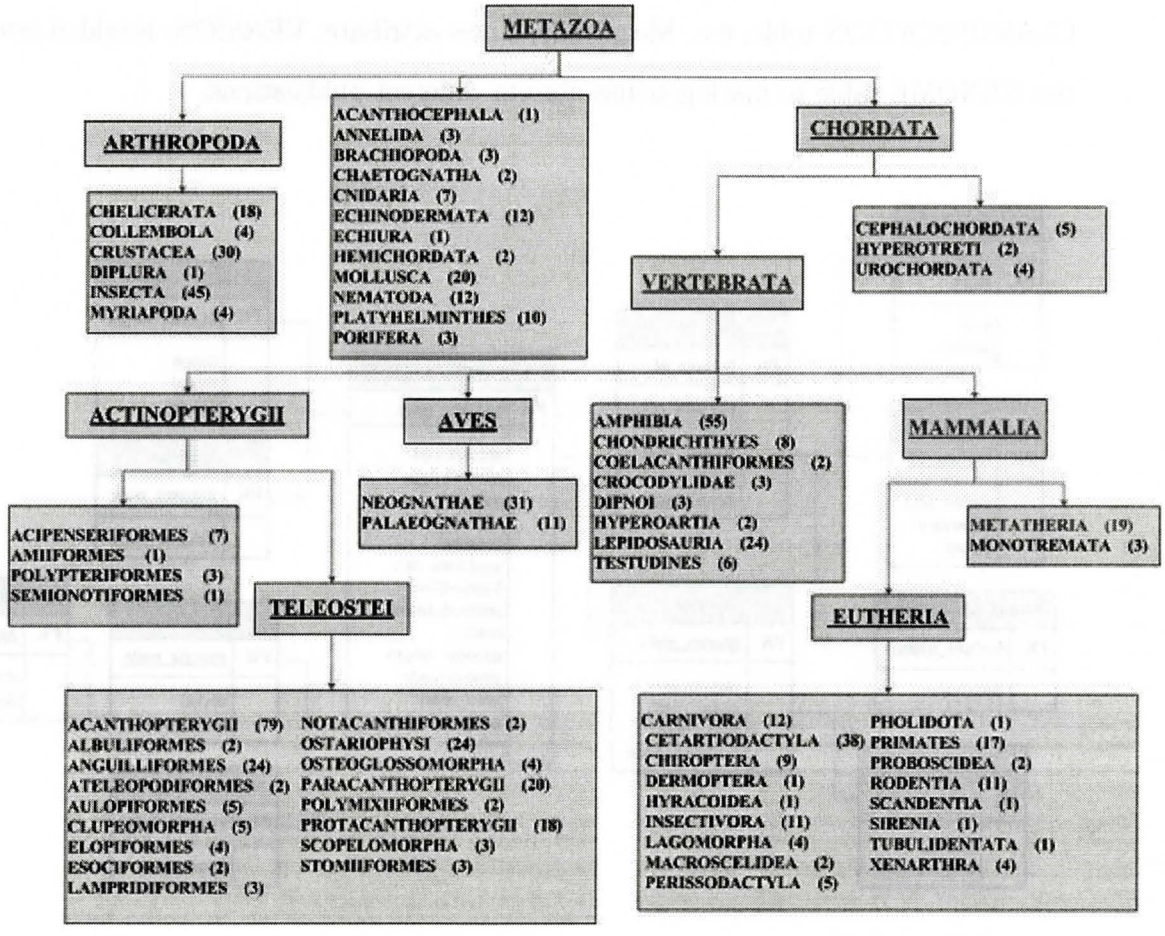

Figure 2.3: The hierarchical structure of taxa in CLASSIFICATION , Nov 2005

In addition, some attributes in different tables are deleted, which are not used any more or can be calculated automatically, including the ALIGNMENT_FILE field in the FEATURE table, the A_CONTENT, C_CONTENT, G_CONTENT, T_CONTENT, RNA_A_CONTENT, RNA_C_CONTENT, RNA_G_CONTENT, RNA_T_CONTENT and FINAL fields in the GENOME table, the ALTERNATIVE_PARENT field in the CLASSIFICATION table, etc. Meanwhile, a new attribute, VERSION, is added into the GENOME table to mark genomes for our different publications.

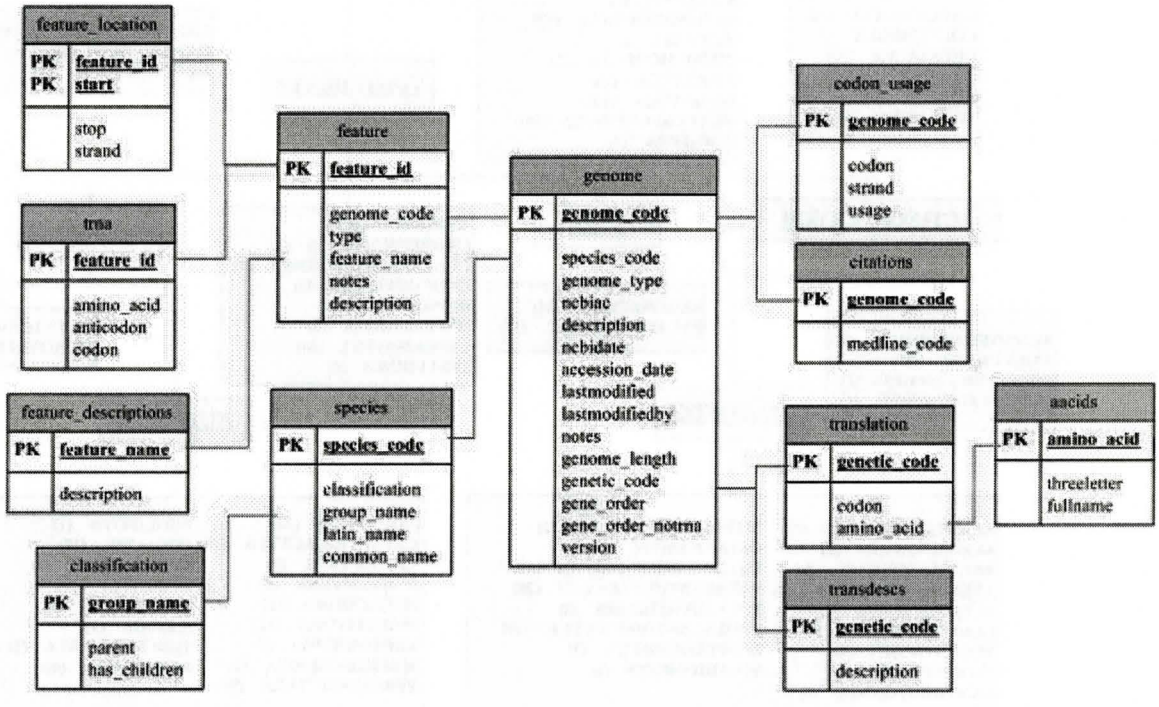

Figure 2.4: The schema of improved OGRe

The mitochondrial gene orders are highly conserved within related species, so they can diagnose the membership of organisms to particular phylogenetic groups. In OGRe, gene order is an intrinsic property of each genome, which is pre-computed and stored in the GENOME table in order to search it quickly. During the update

of OGRe, gene orders are computed automatically after most of tables are reloaded. Besides the attributes in the GENOME table, the original database also includes a GENE_ORDER table and a GENE_ORDER_NOTRNA table to keep gene orders with and without tRNA. However, the two tables only store gene orders and their typical genomes. Each time when OGRe is updated, gene orders of all genomes that are already in OGRe are created again because the features of these genomes may be modified. The gene order of the typical genome might change to another gene order. Hence, it becomes complicated to keep records in the two tables consistent with those in the related attributes in the GENOME table. On the other hand, compared with the GENOME table, the two tables are nearly redundant. Therefore, the two tables are deleted and the typical genome of each gene order is set to be the first genome of all genomes in alphabetical order, which have this gene order. The schema of improved OGRe is shown in Figure2.4.

2.3.4 The Identification of New Genomes

Although it is necessary that OGRe keeps pace with GenBank, it is not reasonable to reload the hundreds of genomes every time when OGRe is updated. In fact, the original parsers implements two steps to update OGRe. First, they find new genomes and then insert them into OGRe. To find new genomes, the original parsers compare the scientific name of each complete mitochondrial genome in GenBank with those already in OGRe. If there is an identical one, the parsers consider the genome as an old one. If not, they will treat it as a new genome and print out its scientific name and accession number. However, sometimes, a record of a genome is deleted while a new one with the same scientific name is added into GenBank. But the parsers can not figure out this kind of changes. Hence, the parsers were modified and accession numbers were included into this process. The new parsers will check both the scientific

names and the accession numbers of genomes because it is nearly impossible that both of them are changed at the same time. Therefore, the genomes downloaded from GenBank are separated into three types: genomes with new scientific names and accession numbers; genomes that have corresponding records in OGRe with different scientific names but the same accession numbers; genomes that have corresponding records with the same scientific names but different accession numbers. Genomes of the first type are thought as new genomes, which can be inserted into OGRe by the improved parsers automatically. The other types of genomes will be examined by hand and then their matching records in OGRe will be modified.

2.3.5 Other Improvements

Although, the genomes downloaded from GenBank are supposed to have complete sequences, the results of retrieve include a few species that miss one or two genes, whose sequences are referred to as 'complete'. Hence, the doubted genomes are compared with other genomes with the same lineages. If the similar genomes have long genes while the doubted genomes have short ones or even don't have genes in the equivalent positions, the doubted genomes are thought to be uncompleted. To notice the problems, a new defined feature, GAP, will be inserted into the gaps in the doubted genomes, whose start and stop positions are equal and configured to be the start point of gaps. Meanwhile, on the website of OGRe, black blocks are painted in diagrams of the uncompleted genomes to remind people. For those genomes whose annotation is slightly incomplete, genes are located by aligning them with similar genomes and added into OGRe.

2.4 The Future Work

The improvement of the update process of OGRe allows new genomes to be added into OGRe automatically. However, those genomes that are already saved in OGRe are not considered. Thus, the next steps can be automatically finding genomes that have been changed in GenBank since they were added into OGRe and automatic modification of their corresponding records in OGRe. Furthermore, we would like to add automatic methods to check for the common errors that we find in GenBank files and avoid incorporating them into OGRe. For example, automatic checking for genes that are expected to be present on the genome but which are not annotated in the GenBank file can be another work for OGRe, which can be solved by aligning closely related species. Similarly, the future work for OGRe includes automatic checking for genes that are longer or shorter than expected in comparison to known genes in species already in the database and checking for genes that are labelled on the wrong strand. Furthermore, the sequence alignments of related species can be saved in OGRe, or the sequences alignments of any selected species can be produced automatically by OGRe. In addition, OGRe can be expanded to the mitochondrial genomes of other groups of eukaryotes and possibly to chloroplast genomes.

Chapter 3

Codon Usage in Mitochondrial Genomes

3.1 Introduction

In mitochondrial genomes, mutation pressure causes a wide variation of base frequencies among species (Foster *et al*, 1997; Singer *et al*, 2000; Urbina *et al*, 2006). Due to the asymmetry of the replication process (discussed in section 1.1), the mutation processes on the two strands are not equivalent. It causes different base frequencies between strands, unequal frequencies of G and C and unequal frequencies of A and U (Bielawski and Gold, 2002; Faith and Pollock 2003; Urbina *et al*, 2006). Besides, the mutation may be context dependent in mitochondria, one of whose signature is that the frequencies of dinucleotides and trinucleotides differ from their expected frequencies if there were no correlation between neighbouring bases.

As discussed in section1.1, codon usage can also be influenced by translational selection. The usual type of selection, translational efficiency, can occur between synonymous codons, if codons that interact more effectively than another with the anticodon are chosen to match tRNA. Its presence is usually demonstrated by determining the codons that are preferred in highly expressed genes and showing that preferred codons have higher frequencies in highly expressed genes than weakly expressed ones. Unfortunately, the corresponding test in mitochondrial genomes is not possible because expression levels have not been measured and it is presumed that

all genes on the mitochondrial genome are essential genes that would have similar expression levels.

Another type of translational selection, translational accuracy, has been demonstrated by showing that preferred codons are more frequent at sites where the amino acid is conserved during evolution (Akashi, 1994; Stoletzki and Eyre-Walker, 2007). It is argued that sites that are critical for protein function have conserved amino acids and that translational accuracy is most important at the critical sites; hence preferred codons should be used at sites that are evolutionary conserved.In contrast, selection for translational efficiency should be equally important at all sites on the gene. The comparison of conserved and variable sites can be done within any gene, and does not require knowledge of the expression level.

Although several methods have been proposed to test the presence of translation selection that might act on the top of mutational effects, none of them seems appropriate for mitochondrial genomes. Calculation of the frequency of optimal codons (Ikemura, 1981) or the codon adaptation index (Sharp and Li, 1987) can only be done if the optimal codons are known, which is not feasible because for, mitochondrial genomes, it is not known which codons are optimal, or even if there are any optimal codons at all. Another standard method is measurement of the effective number of codons, N_c (Wright, 1990). This does not require prior knowledge of the optimal codons. However, a gene might have a low N_c because of mutational bias or because translational selection. Measuring N_c would not be useful to distinguish the factors that affect codon usage. Hence, we develop ways to test for translational selection that are applicable in mitochondria without prior knowledge of expression levels of genes or optimal codons, and that control effectively for the complex nature of the mutation process that is occurring in mitochondrial genomes, which include comparing conserved and variable sites to test translation accuracy.

Aa	Codon	Usage	Aa	Codon	Usage	Aa	Codon	Usage	Aa	Codon	Usage
F	UUU	69	S	UCU	29	Y	UAU	35	C	UGU	5
F	UUC	139	S	UCC	99	Y	UAC	89	C	UGC	17
L	UUA	65	S	UCA	81	*	UAA	4	W	UGA	90
L	UUG	11	S	UCG	7	*	UAG	3	W	UGG	9
L	CUU	65	P	CCU	37	H	CAU	18	R	CGU	6
L	CUC	167	P	CCC	119	H	CAC	79	R	CGC	26
L	CUA	276	P	CCA	52	Q	CAA	82	R	CGA	28
L	CUG	42	P	CCG	7	Q	CAG	8	R	CGG	0
I	AUU	112	T	ACU	50	N	AAU	29	S	AGU	11
I	AUC	196	T	ACC	155	N	AAC	131	S	AGC	37
M	AUA	165	T	ACA	132	K	AAA	84	*	AGA	1
M	AUG	32	T	ACG	10	K	AAG	9	*	AGG	0
V	GUU	22	A	GCU	39	D	GAU	12	G	GGU	16
V	GUC	45	A	GCC	123	D	GAC	51	G	GGC	87
V	GUA	61	A	GCA	79	E	GAA	63	G	GGA	61
V	GUG	8	A	GCG	5	E	GAG	15	G	GGG	19

Table 3.1: The codon usage table of the human mitochondrial genome

3.2 Statistical tests for context-dependent mutation

In this section, the statistical tests are shown to demonstrate the large effect of context-dependent mutation on codon usage in mitochondria. Table3.1 is the codon usage table for the 12 genes on the plus strand of the human mitochondrial genome, which was downloaded from OGRe. We focus on the eight codon families with FFD third positions. In each four-codon family, the usages of four codons are not equal. This indicates that base frequencies at the FFD sites are strongly influenced by the mutational process, because FFD sites are not subject to selection at the protein level. To test if the mutation is context-dependent, a simplest hypothesis is proposed, which is that mutations are independent single-site events and that at there is no selection at FFD sites. Relative frequencies of the four codons in each four-codon family should then be the same. We now show that this is not true.

Let n(XYZ) be the number of occurrences of codon XYZ. Let n(YZ) be the total number of occurrences of each doublet YZ at positions 2 and 3, counting only codons where the third position is FFD, *i.e.* $n(\mathrm{UZ}) = n(\mathrm{CUZ}) + n(\mathrm{GUZ})$; $n(\mathrm{CZ}) = n(\mathrm{UCZ}) + n(\mathrm{CCZ}) + n(\mathrm{ACZ}) + n(\mathrm{GCZ})$; $n(\mathrm{GZ}) = n(\mathrm{CGZ}) + n(\mathrm{GGZ})$. The values $n(\mathrm{YZ})$ form a 3×4 contingency table.The null hypothesis that Z is independent of Y can be tested with a $\chi 2$ test with 6 degrees of freedom (DOF). From the human data in Table 1 we obtain $\chi 2 = 83.57$ ($p < 0.001$), *i.e.* the frequencies of third position bases are definitely not independent of the second position. Differences in the third base frequencies between the individual four-codon families are also studied. The two families CUN and GUN form a 2×4 table with 3 DOF, from which we obtain $\chi 2 = 2.986$ ($p < 0.5$). The four families UCN, CCN, ACN and GCN form a 4×4 table with 9 DOF, from which we obtain $\chi 2 = 14.167$ ($p < 0.1$). Similarly, comparison of CGN and GGN gives $\chi 2 = 8.674$ ($p < 0.05$). These results suggest that there are some differences within the codon families that remain even after the effect of the second base has been accounted for.

Table3.2 shows the results of similar $\chi 2$ tests applied to a set of 148 mammals and 214 ray-finned fish (actinopterygii), which form two comparable but independent monophyletic groups. All species were used that were available in OGRe at the time the project was begun, although the number of species available has increased considerably since then. Each result was classed as not significant if $p > 0.05$, significant if $0.001 < p \leq 0.05$, and highly significant if $p \leq 0.001$. The probabilities of falling into each of these ranges according to the null hypothesis are 0.95, 0.049, and 0.001 respectively. The expected number of species falling into each category is the total number of species multiplied by these probabilities. Table 2 compares the expected numbers with the observed number of species in each significance category. For the UN/CN/GN tests, every species of mammal and fish analyzed falls in the highly sig-

	Not significant $p > 0.05$	Significant $0.001 < p \leq 0.05$	Highly significant $p \leq 0.001$
FISH (214 species)			
Expected No.	203.3	10.5	0.2
UN/CN/GN	0	0	214
CUN/GUN	187	25	2
CGN/GGN	53	101	60
UCN/CCN/ACN/GCN	58	75	81
MAMMALS (148 species)			
Expected No.	140.6	7.3	0.1
UN/CN/GN	0	0	148
CUN/GUN	115	31	2
CGN/GGN	75	58	15
UCN/CCN/ACN/GCN	28	76	44

Table 3.2: Observed and expected numbers of species in each significance category in χ^2 tests.

nificant category. This shows an extremely strong influence of the second position base on the FFD sites. For the CGN/GGN and UCN/CCN/ACN/GCN tests, there are far more species in the significant and highly significant categories than expected. There are also somewhat more species than expected in the significant category for the CUN/CGN tests, but the effect is weaker than for the other cases.

Context-dependent mutational processes create correlations between neighbouring bases. They can be measured by the ratio $R(YZ) = f_{YZ}/(f_Y f_Z)$, where f_{YZ} is the frequency of dinucleotide YZ in the second and third positions, again only four-codon families are considered, and f_Y and f_Z are the frequencies of the individual bases in second and FFD third positions.Similarly, the ratio $R(ZX) = f_{ZX}/(f_Z f_X)$ measures correlations between the FFD base Z and the base X at the first position of the following codon. The results of these ratios are listed in Table3.3.They are far from 1, which indicates non-independence of neighboring sites. The largest ratio is

R(GG) in both data sets and both tables, while the smallest is R(CG) in all cases. The frequency ratios for mammals and fish are correlated with one another: Pearson correlation coefficient $r = 0.90$ for the YZ ratios, and $r = 0.89$ for the ZX ratios. There is also a correlation between the YZ and ZX ratios (using only the 12 dinucleotides for which R(YZ) can be measured). For the fish data from the top and bottom sections of Table3.3, $r = 0.74$, and for the mammals, $r = 0.57$. The results suggest that there are context-dependent mutational effects influencing dinucleotide positions in a similar way in both mammals and fish and acting in a similar way in both reading frames. Similar outcome of the YZ frame were observed by the study of relative dinucleotide frequencies in vertebrate mitochondrial genomes (Shioiri and Takahata, 2001), which considered all pairs, regardless of reading frame or whether the sequence is coding or non-coding. We would expect that context dependent effects on mutation act at the DNA level, and are therefore independent of the frame. However, the observed dinucleotide frequencies also depend on selection at non-synonymous sites, which acts differently at different codon positions. As noted above, there is a reasonable correlation between the frequency ratios in the YZ and ZX frame, but there is no reason why they have to be exactly the same.

In addition, the conditional probabilities are calculated, which is $P(Z|Y) = f_{YZ}/f_Y$ that the FFD base is Z given that the second position base is Y. As an example, Figure3.1 shows $P(\mathrm{U}|\mathrm{Y})$ as a function of f_U for all the fish species, in which the solid line is the equality line ($y = x$) and the dashed lines are the linear regressions to the data sets. In Figure3.1, f_U varies from 0.1 to 0.4 in the fish genomes. The conditional probabilities follow linear trends as a function of f_U. It is shown that $P(\mathrm{U}|\mathrm{U})$ is consistently higher than f_U, while $P(\mathrm{U}|\mathrm{G})$ is consistently lower, and $P(\mathrm{U}|\mathrm{C})$ is slightly less than f_U. This agrees with Table3.3, in that $R(\mathrm{UU}) = 1.250$, $R(\mathrm{GU}) = 0.605$ and $R(\mathrm{CU}) = 0.939$. Figure3.1 indicates that the direction of the

R(YZ)		Mammals				Fish			
		Z				Z			
		U	C	A	G	U	C	A	G
Y	U	0.939	0.743	1.136	1.433	1.250	0.756	1.030	1.274
	C	1.101	1.163	0.906	0.552	0.939	1.205	0.938	0.554
	G	0.763	1.005	1.027	1.654	0.605	0.878	1.145	1.891
R(ZX)		X				X			
		U	C	A	G	U	C	A	G
Z	U	0.855	0.994	1.206	0.856	0.933	0.918	1.096	1.049
	C	1.082	1.363	0.945	0.546	1.162	1.371	0.849	0.609
	A	0.996	0.797	0.974	1.293	0.907	0.739	1.135	1.228
	G	1.115	0.873	0.776	1.369	0.911	0.839	0.758	1.499

Table 3.3: Dinucleotide frequency ratios

context dependent mutational effects seems to be the same in all these species, *i.e.* there is a consistent preference for certain dinucleotides, despite the variation of overall single nucleotide frequencies. The graphs of the conditional probabilities for other base combinations also follow linear trends in both mammals and fish.

It is known that one process that causes context-dependent mutational rates is the 'CpG effect' that occurs in vertebrate nuclear genomes. The C in this context tends to be methylated, and undergoes a rapid deamination to T. This leads to a decrease in CG dinucleotide frequency, and an increase in TG and CA. Karlin and Mrazek (1997) find that R(CG) is very low in a large range of eukaryotic genomes. They also point out that CG is low in mitochondrial genomes, but that this cannot be due to methylation. If a specific process, like the CpG effect is thought to be the dominant context dependent effect in a genome, it is possible to use a mutation rate model specifically for this, and to calculate the way this affects dinucleotide frequencies and sequence evolution over time (Arndt and Hwa, 2005). However, in mitochondria, we do not have accurate information to propose a detailed rate model, and presumably there are many context dependent effects, not just one single process. Here, we use

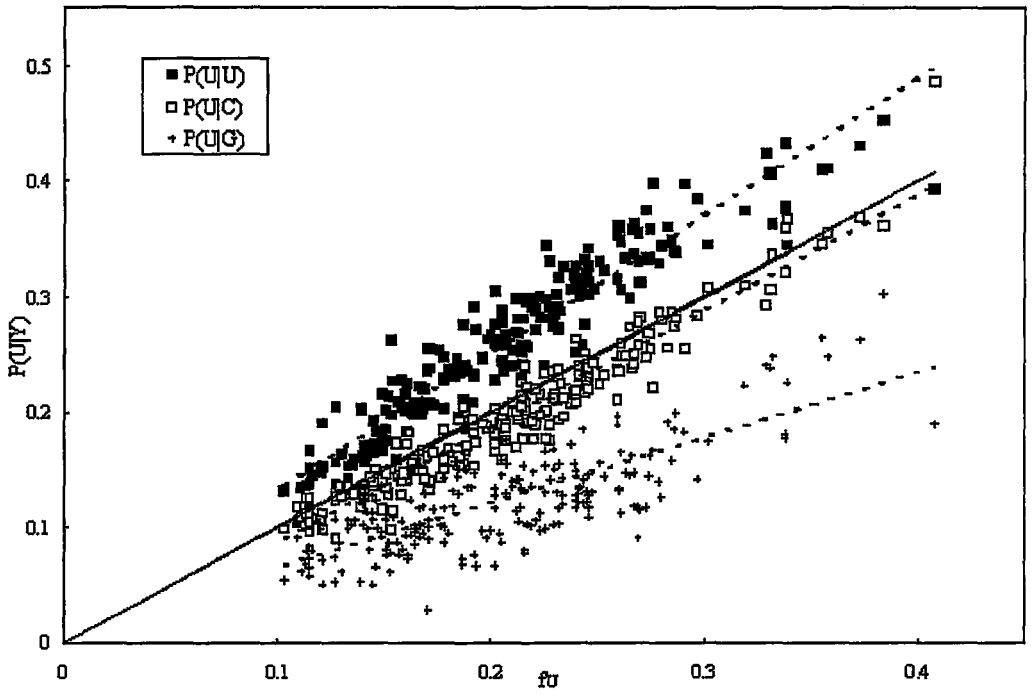

Figure 3.1: Conditional probabilities of that FFD base is U, given that the second position base is either U, C or G

statistical models that predict codon frequencies under a variety of assumptions about mutation and selection, but do not use a specific model for the rate matrix.

3.3 Likelihood-based tests for context-dependent mutation and translational selection

The statistic tests in section 2 clearly demonstrate the major influence of context-dependent mutation; however, the possibility remains that translational selection also has an effect. Hence, the presence of translational selection should be tested in a way that controls for context-dependent mutation. A first point to note is that YZ correlations are likely to have a much greater effect on codon usage than ZX correlations, even if context-dependent mutation rates are equivalent in all frames. This is because, for any given codon family, the second base is always the same, whereas the following first position base is variable. The net influence of the following base is the sum of the different positive and negative influences that would arise from the four possible nucleotides at position X; therefore the effect will be averaged out somewhat. So we will control for Y before any other factor. This will be done by comparing codons for amino acids with the same Y. The four-codon families for Ser, Pro, Thr and Ala (SPTA) all have Y = C. The codonsfor Tyr, His, Asn and Asp (YHND) all have Y = A, and have only U or C at third position. These two sets will be tested separately.

Table3.2 already showed significant differences among the frequencies in SPTA codon families for many species; therefore codon usage is apparently dependent on the amino acid. This may be caused by selection from codon-anticodon matching that occurred in a different way for the tRNAs for different amino acids. However, it is possible that this is merely due to the influence of the nucleotide at position X. The

frequencies of the four nucleotides at position X are determined by the amino acid sequences of the proteins. If amino acid sequences were random, the X frequencies would be the same for all four codon families; therefore, the influence of X would be the same in each case and this would not explain the difference in codon frequencies among the four families. However, amino acid sequences are determined by the function of the protein. Therefore it is possible that the frequencies of the nucleotide at position X following a Ser codon are different from those following a Pro codon, for example. In this case, the influence of X could cause an apparent difference in codon usage between codons for different amino acids.

As discussed in the introduction, selection for translational accuracy (but not efficiency) can lead to an increase in frequency of preferred codons at conserved sites in comparison to variable sites in the same gene. Here, to test its presence for the current mitochondrial sequences, we use a method to compare conserved and variable sites without knowing which codons, if any, are preferred. Furthermore, we note that base frequencies at FFD sites vary systematically with the position along the genome. Base frequencies in each gene depend on D_{ssH}, the amount of time that the gene spends in a single stranded state during the replication of the mitochondrial genome. We will also include models that account for this. Both the translational accuracy factor and the D_{ssH} factor might lead to spurious apparent dependence of the codon usage on the amino acid. For example, one amino acid might have a higher frequency than another in sites that are conserved, or in genes with low D_{ssH}.

We carried out the following series of tests in order to investigate all these possible effects simultaneously. We chose 19 representative mammals, one from each order, and 21 representative fish, one from each major taxon defined in OGRe, making 40 species in all. These species are listed in the caption to Table 3.4. Amino acid sequences of the 12 plus-strand genes were aligned for the mammals and fish separately. A

site was classed as conserved if the same amino acid was present in all 19 mammals or all 21 fish. Otherwise it was classed as variable. To test for the variability of base frequencies along the genome, we divided genes into two groups according to D_{ssH}. Low D_{ssH} genes are COI, COII, ATP8, ATP6, COIII, and NDIII. High D_{ssH} genes are ND4L, ND4, ND1, ND5, ND2, and CytB. The positions of the genes on the genome are the same for these vertebrate species, therefore the classification into the low and high D_{ssH} categories is consistent. In fact, base frequencies at FFD sites vary smoothly with D_{ssH} (Urbina *et al,* 2006) and there is no sharp dividing line between low and high D_{ssH}. However, the two category model below is the simplest model that includes the effect, and it allows a statistical test for the presence of variable base frequencies along the genome that can be done in an analogous way to the test for translational accuracy. In the same way, rates of amino acid substitutions vary in a continuous fashion among sites and there is no sharp division between conserved and variable sites. Nevertheless, the two-category model provides a useful test for translational accuracy.

For each of the 40 species, we counted the codon numbers, $n_{abdX}(Z)$. These are the numbers of times that a codon ending with base Z is used for amino acid a at site-type b in genome position d with following first position base X. The index a takes four values – in one test, SPTA, and in a separate test, YHND. The index b takes two values – conserved and variable. The index d takes two values – low D_{ssH} and high D_{ssH}. The index X takes four values – U, C, A and G. For the SPTA test, Z can be U, C, A or G, and for the YHND test, Z can only be U or C.

We consider models that predict the frequencies $P_{abdX}(Z)$ of codons ending in Z for amino acid a at site-type b in genome position d with following base X. The log of the likelihood of observing the data for any one species is

$$lnL = \sum_a \sum_b \sum_d \sum_X \sum_Z (n_{abdX}(X) ln P_{abdX}(Z))$$

We will compare likelihoods of several models of this form. Akaike's Information Criterion (AIC)is a convenient statistical method of model selection that selects models with high likelihoods but penalizes those with unnecessarily large numbers of parameters. It is defined as

$$AIC = 2(-ln\hat{L} + K), \quad (2)$$

where K is the number of free parameters in the model, and the 'hat' denotes that the ML value has been obtained by fitting the data. The preferred model is the one with the smallest AIC. The statistical theory of the AIC is described by Burnham and Anderson (1998) and an example of its use in molecular evolution is given by Higgs *et al.* (2007).

Model 0 is the simplest possible model for the codon frequencies. It is assumed that there is one overall set of base frequencies $P(Z)$, and that $P_{abdX}(Z) = P(Z)$ for all a, b, d and X. The frequencies must satisfy the constraint that $\sum_Z P(Z) = 1$. Therefore the number of free parameters in the model is $K = F$-1, where F is the number of codons in the family ($F = 2$ for YHND and 4 for SPTA). The ML parameter values are

$$P(Z) = n(Z)/(\sum_{Z'} n(Z')) \ , \qquad \text{where } n(Z) = \sum_a \sum_b \sum_d \sum_X n_{abdX}(Z) \ . \quad (3)$$

Model A assumes that base frequencies depend on the amino acid a but not on the other quantities, *i.e.* $P_{abdX}(Z) = P_a(Z)$ for all b, d and X. The number of parameters is $K = 4(F$-1), as there are four amino acids in each of the test groups. The ML parameters are

$$P_a(Z) = n_a(Z)/(\sum_{Z'} n_a(Z')) \ , \qquad \text{where } n_a(Z) = \sum_b \sum_d \sum_X n_{abdX}(Z) \ . \quad (4)$$

In an analogous way, we can define three other single-factor models, B, D and X, where the frequencies depend on only b, only d and only X, respectively. The numbers

of parameters for these models are $K = 2(F\text{-}1)$, $2(F\text{-}1)$ and $4(F\text{-}1)$, respectively. The ML parameters are given by formulae equivalent to Equation (4).

Next, we define two-factor models AB, AD, AX, BD, BX, and DX where the frequencies depend on two factors only. For example, in model AB, $P_{abdX}(Z) = P_{ab}(Z)$ for all d and X. Thenumber of parameters is K =$8(F\text{-}1)$, and the ML parameters are

$$P_{ab}(Z) = n_{ab}(Z)/(\sum_{Z'} n_{ab}(Z')) \ , \qquad \text{where } n_{ab}(Z) = \sum_{d}\sum_{X} n_{abdX}(Z) \ . \ (5)$$

The other two-factor models are defined in a similar way. Additionally there are four three-factor models in which frequencies depend on three of the four factors. The ML parameters are defined by obvious analogy to the two-factor models above. Finally, there is an exact model E, where the probabilities depend on all four factors. This is an exact fit of the data because the number of parameters is equal to the number of independent quantities in the data: $K = 64(F\text{-}1)$. The ML parameters are

$$P_{abdX}(Z) = n_{abdX}(Z)/(\sum_{Z'} n_{abdX}(Z')) \ . \ (6)$$

Table3.4 shows the results of fitting these models to the 40 species. We quote ΔAIC, which is the change in AIC with respect to model 0. A ΔAIC of order 1 denotes a slight preference for one model over another, and a ΔAIC of order 10 is usually considered sufficient to rule out the less well fitting model. The results from the human mitochondrial genome are given as an example. For both SPTA and YHND, model X has the lowest AIC, followed by model DX. Factor X has a larger influence than any of the other factors, as seen by the large negative ΔAIC for model X. Models A and D also have negative ΔAIC, indicating some apparent influence of factors A and D. However, when these factors are combined with X, there is no significant improvement over factor X alone, *i.e.* models AX and DX have a higher AIC than model X. All the three-factor models and model E have positive ΔAIC, indicating that these models are over-fitting the data.

Rather than quote AIC values for every species separately, we have summarized the results on the 40 species in several ways. Table3.4 shows the average ΔAIC for each model. For each species, ΔAIC is measured relative to the AIC for model 0 for that species, and the mean of the ΔAIC is then calculated. For both SPTA and YHND, model X has the lowest average ΔAIC, and model DX has the second lowest. The human example is thus typical of the majority of the species.However, the average values mask considerable variation among the species. The third pair of columns in Table3.4 gives the number of species for which $\Delta\text{AIC} < 0$ for each model (*i.e* the number of species for which the model is an improvement over the null assumption). Once again, models X and DX score highest by this criterion. Inclusion of factor X gives an improvement for almost all cases. The three-factor models and model E perform poorly because they over-fit the data for most species.

The fourth pair of columns compares the single-factor models with model 0 by showing the number of species for which each model has the lowest AIC of these five models. Clearly X is the dominant factor for the majority of species, there are a few species for which factors A and D are dominant, and there are no species for which factor B is dominant. The fifth pair of columns gives the number of species for which each model is best when all models are considered. Once again X and DX score highly, although AX scores almost as highly as DX. In roughly half the cases, a single-factor model is best, and in roughly half the cases a two-factor model is best. There is only one case where a three factor model is best. Finally, we quote the number of species for which each of the four factors is included in the best model, either as a single factor or as part of a two- or three-factor model.

There are several clear conclusions from Table3.4. Factor X has a large effect in almost all species. This shows that context dependent mutation causes correlations

Model	ΔAIC for *Homo sapiens*		average ΔAIC		# of species for which ΔAIC < 0		# of species for which this model is the best of 0 and single-factor models		# of species for which this model is the best of all models		# of species for which this factor is included in the best model	
	SPTA	YHND	SPTA	YHND	SPTA	YHND	SPTA	YHND	SPTA	YHND	SPTA	YHND
0	0.00	0.00	0.00	0.00	—	—	0	3	0	2	—	—
A	-2.02	0.67	-6.96	-2.38	28	21	6	9	3	2	12	15
B	5.26	1.92	2.78	1.13	6	6	0	0	0	0	2	7
D	-3.06	0.16	-2.26	-0.65	27	11	2	5	0	3	8	14
X	-19.08	-9.34	-29.61	-7.51	38	33	32	23	21	8	34	27
AB	12.45	-2.60	3.97	-2.59	15	22	—	—	1	3	—	—
AD	4.30	0.76	-1.62	-0.53	22	16	—	—	2	2	—	—
AX	-6.49	5.54	-19.12	-2.03	33	16	—	—	6	7	—	—
BD	6.85	4.13	2.77	1.11	14	8	—	—	0	0	—	—
BX	-2.24	-3.98	-18.15	-3.71	32	25	—	—	1	3	—	—
DX	-17.50	-7.17	-24.09	-5.75	38	28	—	—	6	9	—	—
ABD	31.09	9.97	19.17	5.74	4	11	—	—	0	1	—	—
ABX	34.15	19.51	26.27	11.47	9	8	—	—	0	0	—	—
ADX	36.23	17.58	21.85	9.95	9	10	—	—	0	0	—	—
BDX	9.54	0.51	-2.48	1.28	20	17	—	—	0	0	—	—
E	141.18	44.94	118.19	34.96	0	1	—	—	0	0	—	—

Table 3.4: Results of the model selection process applied to 40 representative species of mammals and fish: *Homo sapiens*; *Canis familiaris*; *Sus scrofa*; *Artibeus jamaicensis*; *Cynocephalus variegates*; *Procavia capensis*; *Erinaceus europaeus*; *Oryctolagus cuniculus*; *Elephantulus sp. VB001*; *Equus caballus*; *Manis tetradactyla*; *Elephas maximus*; *Mus musculus*; *Tupaia belangeri*; *Dugong dugon*; *Orycteropus afer*; *Dasypus novemcinctus*; *Didelphis virginiana*; *Ornithorhynchus anatinus*. *Acipenser stellatus*; *Amia calva*; *Polypterus ornatipinnis*; *Lepisosteus oculatus*; *Takifugu rubripes*; *Albula glossodonta*; *Anguilla anguilla*; *Ateleopus japonicus*; *Aulopus japonicus*; *Engraulis japonicus*; *Elops hawaiensis*; *Dallia pectoralis*; *Lampris guttatus*; *Notacanthus chemnitzi*; *Cyprinus carpio*; *Osteoglossum bicirrhosum*; *Gadus morhua*; *Polymixia japonica*; *Salmo salar*; *Diaphus splendidus*; *Gonostoma gracile*

between third position bases and following first position bases, and hence has a major influence on codon usage in mitochondrial genomes. Factor B has very little effect. There is no example where model B is best and only a small number of cases where factor B is included in the best model as part of a two- or three-factor model. We therefore conclude that selection for translational accuracy does not have a significant influence on codon usage in most mitochondrial genomes. Although factors A and D have less importance than factor X, there are still several examples where these factors are included in the best model. The presence of factor D is understandable. It has already been demonstrated that that base frequencies vary systematically along the genome, and this can be understood in terms of the position of the genes with respect to the origin of replication. These results show that the effect of variability along the genome is sufficiently strong to merit theinclusion of extra parameters in the model in at least some species.

The presence of factor A in the best model for several species is difficult to interpret. Ostensibly, this says that codon usage differs among amino acids, but it does not say why. If we looked at model A alone, we would conclude that this factor is important in more than half the examples (28 and 21 for SPTA and YHND). However, a spurious dependence on A can arise because of a correlation between A and other factors. After accounting for the other factors, we find that factor A remains in the best model in many fewer cases (12 and 15 for SPTA and YHND).Thus, to some extent, the apparent dependence on A is due to the real dependence on X and D, both of which are mutational effects. However, there are species for which models AX or AD are better than X or D alone. Can this be explained by translational selection? Although the lack of importance of factor B rules out translational accuracy, it does not rule out translational efficiency.

In many bacteria, there are clear differences in codon usage among amino acids that correlate with the anticodons of the tRNAs, and the bias is stronger in highly expressed genes. This is an obvious indication of selection for efficient codon-anticodon matching. However, the anticodons of the mitochondrial tRNAs in the SPTA set all have wobble position U and middle position G and differ only in the third anticodon position (which pairs with the first codon position). Similarly the mitochondrial tRNAs for YHND all have wobble position G and middle position U and differ only at the third anticodon position. Thus the only way that selection for codon-anticodon matching could explain the relevance of factor A is if the selection on the third codon base caused by the wobble base were dependent on the bases at the first codon and third anticodon position. However, if this kind of selection was to be responsible for differences in codon usage among these codon families, it would have to operate in a very subtle way that is not simply a function of the wobble base. We therefore feel that this is an unlikely explanation.

With the above in mind, it seems that a more plausible explanation for the relevance of factorA is, once again, context-dependent mutation. The models above control for the middle position base (by separating amino acids into groups with the same Y) and for the following first-position base (by including X in the model), but they do not consider longer range correlations. If the mutation process depends on nearest neighbours, this automatically sets up correlations between a site and its neighbours, and the neighbours of the neighbours, and so on. Thus, it is quite possible to create a significant correlation between a third position base and the first position base in the same codon. This is exactly what is described by model A.

3.4 Is there any detectable influence of the wobble base on codon usage?

From all the previous considerations, it is likely that the influence of the wobble base of anticodons on codon usage in mitochondria is weak. In this section, we set out to test this by looking for codons of Met where an effect should show up, if there were one. Met is coded by two AUR codons (R = A or G) in the vertebrate mitochondrial genetic code. Two-codon families with A and G codons usually have a wobble-U tRNA, e.g. Gln, Lys, Glu and Leu(UUR). However, the Met tRNA has a C at the wobble position in the gene, which is modified to f^5C in the tRNA molecule. This is a relic of the reassignment of the AUA codon from Ile to Met that occurred in most mitochondrial genomes. In genomes where the standard code is used, the wobble position is an unmodified C, which pairs only with the G-ending codon. Modification of this base allows it to pair with both codons after the codon reassignment. We define $g_G = n(G)/(n(A) + n(G))$, the relative frequency of G-ending codons with respect to the sum of A and G codons. We compare Met codons with Leu(UUR) codons because they have the same second base. Figure3.2 shows that g_G is larger for Met than for Leu(UUR) for all 326 species of actinopterygian fish in the current OGRe database. Thus, we conclude that the f^5C base exerts a preference for the G codon, even though it is capable of translating the A codon too. Translational selection is elevating the frequency of G-ending codons for Met due to the unusual nature of the tRNA-Met. This was pointed out previously by Xuhia Xia (Xia, 2005).This is the only example in vertebrate mitochondria we are convinced that there is a significant influence of the wobble position on codon usage.

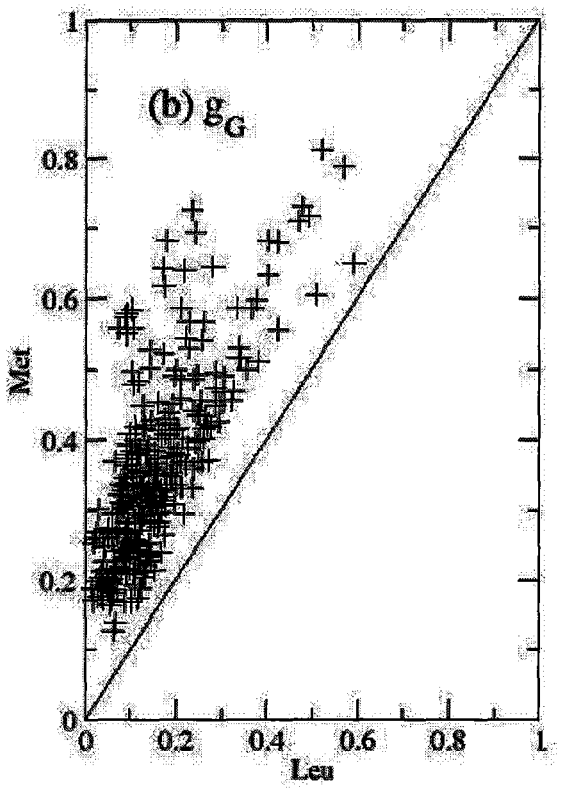

Figure 3.2: Relative frequencies of third position G bases in Met codons versus Leu(UUR) codons. Data are from 326 current fish genomes.

3.5 Discussion

Mutation is a major force in mitochondrial genomes because of large mutation rates and rates of evolution of animal mitochondrial genes are usually much faster than nuclear genes, presumably, because of the large mutation rate. It is therefore not surprising that mutation should be dominant in determining codon usage in mitochondria. The mutation process in mitochondrial genomes is complex: rates are different on the two strands (as is apparent from simple comparison of the base frequencies between strands) and are context dependent (as is apparent from the statistical analysis above).

In Section3.2, it is shown that the same dinucleotides seem to be preferred or avoided in mammals as in fish, and that the preference or avoidance of a dinucleotide seems to be consistent, even though the mean single nucleotide frequencies vary tremendously among species in each of these groups. This suggests that the processes causing the context dependent biases may be operating similarly in all these species. Experimental evidence will be required to determine exactly what the chemical processes are that cause these biases.

Although the presence of translational selection in rapidly growing microorganisms such as yeast and many bacteria has long been recognized, it has been much more difficult to distinguish between selection and mutation effects in metazoa (Duret, 2002), where selection may be weaker and where differences in base composition occur within the genome. Detection of a relationship between expression level and codon usage is one of the principal ways of showing the existence of translation selection. In mitochondria, it is not straightforward because there is no easy comparison of high- and low-expression genes. However, the test for translational accuracy used in Section 3.3 does not require this, and we showed that this can be done carefully in a way that controls for the context-dependent mutation. In section3.3, we found that there

was no evidence for translational accuracy selection in almost all the mitochondrial genomes considered. Evidence has previously been given for the importance of translational accuracy in some genomes other than mitochondria (Akashi, 1994; Stoletzki and Eyre-Walker, 2007). The analysis in these papers did not control for context dependent mutation, which is surely a factor in many genomes in addition to mitochondria. Therefore, it would be interesting to know if the evidence for translational accuracy selection would stand if context dependent mutation were also allowed for in the other genomes.

Selection for translational efficiency could, in principle, lead to selection of codons that best match the limited set of tRNAs still present on mitochondrial genomes. The rule that the wobble position is U for four-codon families, U for A+G families, and G for U+C families applies for almost all codon families in almost all the animal phyla. It is difficult to explicitly test for this kind of selection because it could act in the same way on different codon families, and therefore not lead to an observable difference in codon usage between families. In section3.4, we considered a case to test the influence of the wobble base. The AUA codon, which is Ile in the canonical code, is reassigned to Met in most metazoan mitochondrial genomes. This reassignment cause that the tRNA-Met can translate both AUA and AUG (Sengupta *et al.* 2007). Before the codon reassignment, the tRNA-Met has wobble-C. We might expect that this would switch to U when the genetic code changes, like tRNA-Trp, whose wobble base is modified from C to U when the UGA codon is reassigned from stop to Trp. However, in all vertebrate mitochondria, this position remains as C in the tRNA gene sequence. In a few species, we know from experiment that this C is post-transcriptionally modified to f^5C, and it is presumed that this modification occurs in all vertebrates. In Figure3.2, we showed that there is a preference for the G-ending codon for Met. The f^5C is doing the job of translating both codons, but it

is doing less well than would a wobble-U tRNA. This raises the question of why the C does not simply mutate to a U in the gene sequence, making the f^5C modification unnecessary. It is likely that this has something to do with the special role of AUG as an initiator codon. In bacteria, there are distinct initiator and elongator tRNA-Met genes, both with C at the wobble position. This is also true in some non-metazoan mitochondria where AUG is the only Met codon (the case of fungi is discussed in detail in the supplementary information of Sengupta *et al.* 2007). The single tRNA-Met in metazoan mitochondria must be adapted to do both jobs. We suggest (although we have no evidence) that there is a constraint that a C (or modified C) is required at the wobble position specifically when the tRNA acts as an initiator, and that a wobble-U tRNA would not work as an initiator.

We conclude that in the case of Met, there is evidence for weak selection preferring codons that match the anticodons. In contrast to bacteria, where there are many species in which translational selection has an important influence on codon usage, in mitochondria, codon usage patterns seem to be determined principally by complex context-dependent mutational effects.

Chapter 4

The Phylogeny of Birds

4.1 Introduction

Phylogenetics is the study of the history, origin and evolution of related organisms, in particular, is used to find out the ancestral relationships among living or extinct species. The fundamental idea of phylogenetic is that the longer time since two organisms diverged from a common ancestor, the more dissimilar those organisms will be. Therefore by measuring the degree of dissimilarity between organisms we can hope to place them on a tree that shows the evolutionary relationships between them. The tree is usually called a phylogeny. For a long time, phylogenies have been created using morphological characters of species. Since the late 20^{th} century, biochemistry has been another important source of data for phylogenetic studies because the data can be quantified more easily. It has generated enormous numbers of nucleic acids and protein sequences. In particular, rRNA sequences have been widely used, as they are found in all genomes, including those of organelles. To obtain the relationships among organisms, various approaches have been developed to analyse the molecular information. It is thought that the difference between two sequences mirrors their evolutionary distance. In other words, if two organisms have fewer sequence differences, they are more closely related. Therefore, the approach compares the DNA or protein sequences from different organisms and estimates the evolutionary relationships based on their difference.

The phylogeny of a set of organisms is usually expressed as a tree-like diagram that shows the evolutionary relationship between the species. In a phylogenetic tree, leaf nodes represent extant (or sometimes extinct) organisms and the inner nodes represent their ancestors. The positions of nodes represent the relationships among species, in other words, close branches stand for closely related species. Sometimes, the lengths of branches are proportional to the time since the divergence of species, and sometimes the branch lengths represent the number of evolutionary changes. The phylogenetic tree can be rooted or unrooted. A rooted tree is a directed tree with one common ancestor of all leaf nodes, and an unrooted tree is an undirected tree without assumptions about the common ancestor.

The most common types of approach that are used to reconstruct phylogenetic trees are distance methods, maximum parsimony, maximum likelihood and Bayesian inference (Zwickl and Hillis, 2002). Distance methods such as neighbor joining and UPGMA measure genetic distances from multiple sequence alignments and then build a tree to match the difference. These are the simplest and fastest methods and date back to Cavalli-Sforza and Edwards (1967) and by Fitch and Margoliash (1967). Maximum parsimony evaluates all possible trees and chooses the tree with the minimum number of evolutionary changes (Felsenstein, 2004). This is perhaps the most popular method for the reconstruction of phylogenetic trees. Maximum likelihood is a popular statistical method to find values of parameters of probability models and selects those parameters that make their corresponding likelihood a maximum. Bayesian inference assumes prior distributions of related parameters and calculates the posterior distribution of the parameters using the real data.

The PHASE package has previously been developed in our group. It contains RNA specific methods to generate phylogenetic trees of species. These RNA specific methods have previously been shown to give useful results in mammalian phylogenies

using RNA sequences from mitochondria (Jow *et al*, 2002; Hudelot *et al*, 2003). Therefore, we wished to apply these methods to bird sequences using a large set of newly available sequences sent to us by Dr. van Tuinen of the department of Biology and Marine Biology of University of North Carolina Wilmington.

4.2 Phylogenetic relationships among modern birds

Although the higher level relationships among the major avian clades remains unclear, the base of modern birds, which are classified as a taxon Neornithes, is widely accepted. Neornithes is divided into two basal clades, Paleognathae ("Old Jaws") and Neognathae ("New Jaws"). Paleognathae usually have long necks, long legs and are good at running rather than flight. This group contains two subgroups: one is Ratitae, which is represented by the large, cursorial and flightless birds such as ostriches, emus and rheas; the other one is Tinamiformes, which is one of the oldest groups of birds. Neognathae includes all other modern birds. It consists of two sister clades, Galloanserae and Neoaves. Galloanserae contains Galliformes, which is landfowl like pheasants and chicken, and Anseriformes, which is waterfowl, including screamers, geese, ducks, swans and allies. Neoaves contains all the other neognaths. The three level relationships are obtained from morphological and molecular studies (Huxley, 1867; Cracraft, 1981; Sibley and Ahlquist, 1990; van Tuinen *et al*, 1998; Groth and Barrowclough, 1999; van Tuinen *et al*, 2000; Livezey and Zusi, 2001; Mayr and Clarke, 2003; García Moreno *et al*, 2003), which are shown in Figure 4.2. It remain unchallenged that Tinamiformes, Ratites, Anseriformes and Galliformes are monophyletic, *i.e.* each of them contains all species that evolved from one common ancestor (Mayr, 2000; Harrison *et al*, 2004; Dyke and van Tuinen, 2004; Cracraft *et al*, 2004; Livezey and Zusi, 2007).

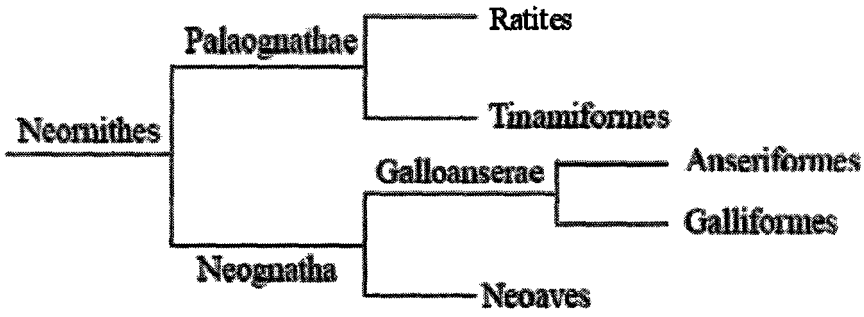

Figure 4.1: The basal relationships of modern birds (Cracraft *et al*, 2004)

Within Neoaves, many of the traditional orders defined by morphological methods are also found to be monophyletic in molecular studies. For example, the monophyly of Charadriiformes (gulls, plovers and allies) has been proved by several studies (*e.g.* Dyke and van Tuinen, 2004; Livezey and Zusi, 2007) although Sibley and Ahlquist (1990) believed that shorebirds belong to Ciconiiformes (storks and allies). Podicipediformes (grebes), Psittaciformes (parrots and allies) and Passeriformes (perching birds) are thought to be monophyletic with little doubt (Dyke and van Tuinen, 2004; Cracraft *et al*, 2004). The monophyly of traditional Piciformes (woodpeckers and allies) was supported by several studies (Raikow and Cracraft, 1983; Johansson and Ericson, 2003) although it was questioned by Olson (1983). However, there are several orders where molecular studies indicate that the traditionally defined orders are not monophyletic and a classification of the names of the orders will be required in future when this becomes clear. For example, Pelecaniformes (pelians and allies) are polyphyletic, in other words, the order contains species evolved from different ancestors (van Tuinen *et al*, 2000; Cracraft *et al*, 2004). Ciconiiformes are also thought to be polyphyletic (van Tuinen *et al*, 2001; Fain and Houde, 2004) although Livezey and Zusi (2007) challenged the opinion. As well, Gruiformes (Cranes and allies) and

Falconiformes (falcons, eagles, hawks and allies) are found to be polyphyletic (Fain and Houde, 2004; Livezey and Zusi, 2007).

In this study, mitochondrial ribosomal RNAs (rRNAs) were used to analyze the phylogenetic relationships within modern birds, which are components of mitochondrial ribosomes and play central roles in catalytic activities (Alberts, 2002). The rRNA gene sequences of 159 avian species were chosen to test existing hypotheses or to find new relationships among them.

4.3 Data and Methods

In this study, nucleotide sequences of mitochondrial 12s and 16s rRNA genes of 159 avian species were used, which were provided by Dr. van Tuinen of University of North Carolina Wilmington. The 159 species are listed in Table4.1. multiple alignments were performed separately for the two genes. The RNA secondary structure was then added to each alignment manually by Dr. Higgs. The secondary structure was represented by bracket and hyphen notation at the top of the plain alignment. An opening bracket and its matching closing bracket indicated that nucleotides at the two positions form a base-pair in the secondary structure of rRNAs (Gowri-Shankar and Jow, 2006). Hyphens demonstrate the base at the position is unpaired (Gowri-Shankar and Jow, 2006). After the structures were done, the alignments of these genes were concatenated and non conserved pairs and sites with too many gaps were removed. Finally, we obtained an alignment with 2425 nucleotides as the total length.

In this study, we applied two methods with different substitution models to generate the phylogenetic trees of birds. The first method used a single site model for all sites. The second method used the RNA pair model for the paired sites and a single site model for the unpaired sites. The alignments for both cases are the same.

The software package PHASE (Gowri-Shankar and Jow, 2006) was used to generate the unrooted phylogenetic trees and consensus trees. Bayesian posterior probabilities, which are used to test the reliability of trees, were also calculated by PHASE. For the first method, we used the REV (Tavare, 1986) model defined in PHASE and six gamma-distributed rate categories (Yang, 1994). For the second method, we used the REV model for the loops of rRNAs and the RNA7A (Higgs, 2000) model for thte stems of rRNAs. Both of the two models were combined with six gamma-distributed rate categories.

Several runs of the MCMC algorithm were performed for each of the two methods. It was found to be computationally time-consuming to deal with all 159 species, and we were not convinced that the runs had fully reached equilibrium. We are particularly interested in the deep-level relationships between the orders of birds. These are the slowest parts of the tree to equilibrate and the most likely to get trapped in local optima during MCMC runs. To speed up the program and to facilitate the equilibrium of the deep level branches on the tree, we divided the 159 species into clusters, as listed in Table4.1. Considering the monophyly of orders introduced in the previous section, a cluster was defined as a subset of species which appeared as a monophyletic group with 100% posterior probability in all the initial runs of the program. Species that did not cluster repeatably with others were left as individuals (*i.e.* one-species clusters). The results presented here are obtained using MCMC runs with the clusters specified. The same two methods were used as in the initial unconstrained runs. In the runs with specified clusters, the space of possible trees is restricted so that the clusters are always present as monophyletic groups. However, the arrangement of the clusters relative to one another is unrestricted, as is the arrangement of species within each cluster. It was found that this procedure considerably shortened the required time for the MCMC runs and led to much more repeatable results.

The names of the clusters are given in Table4.1. In many cases, the clusters correspond to recognized orders of birds, and the latin name of the order is used for the cluster. For the four polyphyletic orders, Gruiformes, Pelecaniformes, Ciconiiformes and Falconiformes, we divided the order into more than one cluster and we have given an English name to the cluster for convenience. Table4.1 also lists nine-letter abbreviations for the species. In cases where there is a single species in a cluster, this is labelled as 'single' in Table4.1 and the abbreviation is used in the trees in Figures 4.1 and 4.2. These abbreviations are also used to label the species in the small trees in Figures 4.3-4.8 that show relationships between species in the same cluster.

4.4 Results and Discussion

Both methods were run twice with different random starting positions. The two phylogenetic trees obtained for each method are consistent. Figure 4.2 and Figure 4.3 are the consensus trees generated with the first and second methods. Bayesian posterior probabilities are shown only if they are larger than fifty percent.

4.4.1 Relationships among orders and families

The two figures show the same relationships among Tinamous, Ratites, Anseriformes, Galliformes and Neoaves, with 100% as the Bayesian posterior probabilities between Anseriformes and Galliformes, between Galloanserae and Neoaves and between Paleognathae and Neognathae. This supports the opinions discussed in section 4.2. Within Neoaves, many relationships among clusters are different in the two figures and a view of rapid evolution is shown in both figures. The hypothesis that Neoaves is divided into Metaves and Coronaves (Fain and Houde, 2004; Ericson *et*

al, 2006) is not supported in either tree. If internal branches that are not supported at with more than 50

The clade of tropicbirds is linked with the mesite, *Mesitornis unicolor*, with 99.7% in Figure 4.2, while 61.8% in Figure 4.3. However, it is difficult to conclude the mesite is most closely related with tropicbirds because mesites are traditionally included in Gruiformes (Fain and Houde, 2004) or grouped with the clade of grebes and flamingos (Ericson, 2006). In both figures, it is also shown that cormorants and allies are grouped with frigate birds with high probabilities. Here, the cluster Cormorants and allies includes two boobies (Sulidae), one cormorant (Phalacrocoracidae) and one snakebird (Anhingidae). The three families have been generally accepted to be related to one another. The topologies of their phylogenetic trees in the two figures are identical. The two boobies are grouped together with 100%, and then they are allied with the cormorant with 98.8% by the first method and 94.8% by the second method. Finally, the three organisms are linked with the snakebird with 100% posterior probability. Although some studies demonstrated that Sulidae is the sister clade of Pelecanidae (pelicans) or Fregatidae (frigatebirds) (Mayr, 2003; Livezey and Zusi, 2007; Mayr, 2007), the relationships among the three families and frigate birds in the two figures support the opinion that (Phalacrocoracidae + Anhingidae) are sister clades of Sulidae, and then their new clade is grouped with Fregatidae (Mayr, 2003; Cracraft *et al*, 2004; Fain and Houde, 2004; Ericson *et al*, 2006; Mayr, 2007).

In the two figures, although the hornbill, *Tockus nasutus*, and the jacamar, *Galbula pastazae*, are allied together with more than 90%, the relationship is not consistent with the opinion of other studies that Upupiformes (hoopoes and wood-hoopoes) is the sister clade of Bucerotiformes (hornbills) (Livezey and Zusi, 2007; Mayr, 2007). However, in Figure 4.3, it is shown that the hornbill, the jacamar, Piciformes (woodpeckers), the hoopoe and the roller are close related species, which agrees with the

other studies although the relationships among them are not consistent with other hypotheses. In Figure 4.3, the Pelicans are grouped with the Spoonbills with 99.99%, which provides a support to the tree based on mitochondrial sequences (Cracraft *et al*, 2004:Figure27.6).

The two figures show that rails are grouped with the sungrebe, *Heliornis fulica*, with 100%, and cranes are grouped with the trumpeter, *Psophia viridis*, with high probabilities (98.6% in Figure 4.2 and 91.3% in Figure 4.3). Then the two new groups are linked together with about 99% and form a new clade. Here the cluster Cranes includes eight cranes and one limpkin. The cluster Rails contains five rails and one coot. The relationships among cranes, rails, the sungrebe and the trumpeter support the argument that (Gruidae (cranes) +Aramidae (limpkin)) are the sister clade of Psophiidae (trumpeters), Rallidae (rails and crakes) and Heliornithidae (sungrebes) are sister clades, finally, the two new clades are linked together (Livezey, 1998; Cracraft *et al*, 2004; Fain and Houde, 2004; Ericson *et al*, 2006). The two sub-trees of cranes generated by the two methods are the same except some small differences in the corresponding Bayesian posterior probabilities. Figure 4.4 is the sub-tree of the second method. It is clearly shown that the group of all eight cranes are linked with the limpkin with 100%, which supports the sister relationship between Gruidae and Aramidae (Cracraft *et al*, 2004; Mayr, 2007). In Figure 4.4, the sister relationship between the crowned crane, *Balearica pavonina*, and the other seven cranes agrees with the phylogenetic trees in several studies based on different types of data (Krajewski and Dickerman, 1990; Krajewski and Fetzner, 1994; Fain *et al*, 2007). However, the relationships among the other seven cranes are not consistent with these studies, which are thought to be the result of insufficient data. For both of the methods, the sub-trees of rails are different not only in the Bayesian posterior probabilities but also in their topologies. Figure 4.5 is the sub-tree of the second

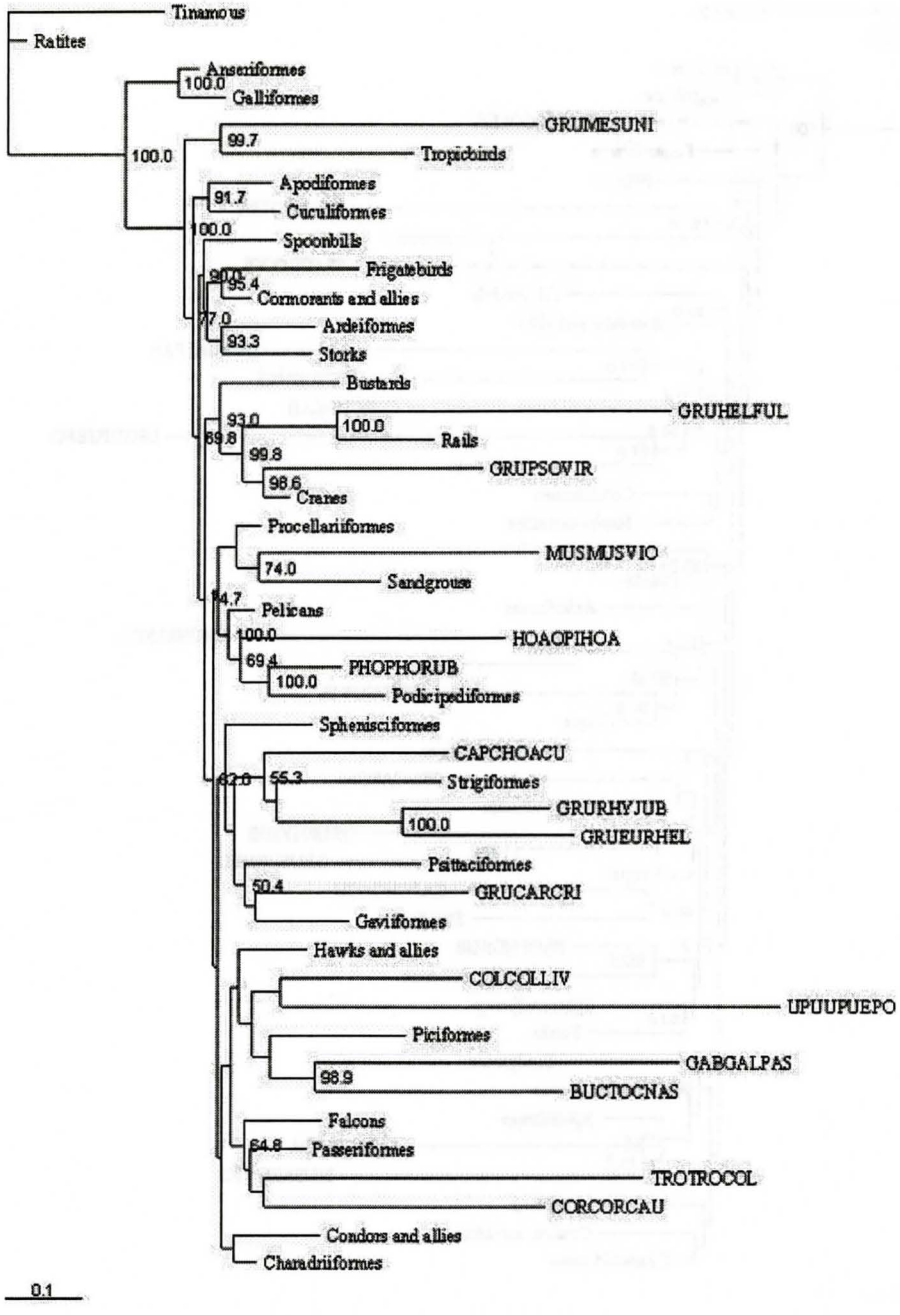

Figure 4.2: The phylogenetic tree of 159 avian species without using the secondary structure of rRNAs

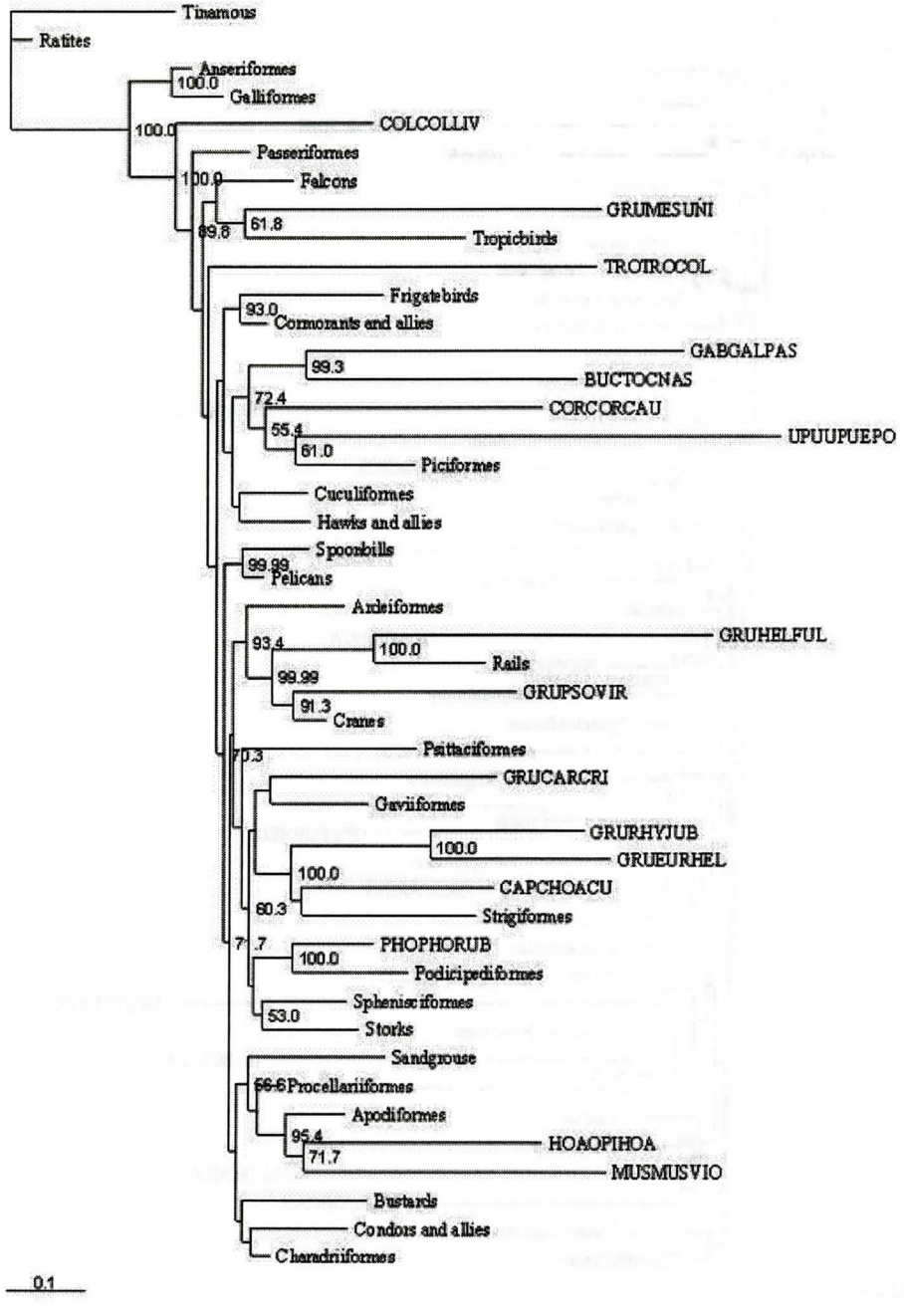

Figure 4.3: The phylogenetic tree of 159 avian species using the secondary structure of rRNAs

method. In the sub-tree of the first method, the six organisms are separated into two clades, one of which is *Rallus longirostris* and *Gallirallus phillipensis*, like in Figure 4.5. For the rest species, *Fulica americana* and *Porzana carolina* are sister taxa with 100%, they are most closely related to *Porphyrio porphyrio* with 61.2%, and then the three species are grouped with *Laterallus melanophaius* with 57.5 %. The clade of *Rallus longirostris* and *Gallirallus phillipensis* agree with the studies of Livezey (1998) and Fain *et al*, (2007). However, only the relationships of the rest rails of the second method are consistent with the study of Fain *et al*, (2007).

The both figures show that the sunbitten, *Eurypyga Helias*, and the kagu, *Rhynochetos jubata*, are grouped together with 100%, which supports the sister relationship between Eurypygidae (Sunbittern) and Rhynochetidae (kagu) (Livezey and Zusi, 2007). Several different independent studies (van Tuinen *et al*, 2001; Chubb 2004; Mayr, 2004; Manegold, 2006; Mayr, 2007) show that Phoenicopteridae (flamingos) is most closely related to Podicipedidae (grebes), whereas Livezey and Zusi (2007) believed that Podicipedidae is the sister clade of Gaviidae (loons) and Phoenicopteridae is the sister clade of Ciconiidae (storks). The sister relationship between grebes and flamingos is supported in the both figures in that the flamingo species, *Phoenicopterus rubber*, is allied to Podicipedidae with 100%. In Figure 4.3, it is shown that Sphenisciformes (penguins) and Storks are the most closely related, which agrees with a recent study (Watanabe *et al*, 2006) although traditional opinion believes the extant closest relative of penguins is Gaviiformes (loons) or Podicipediformes (grebes) (Cracraft, 2004) or Procellariiformes (shearwaters and allies) (van Tuinen *et al*, 2001).

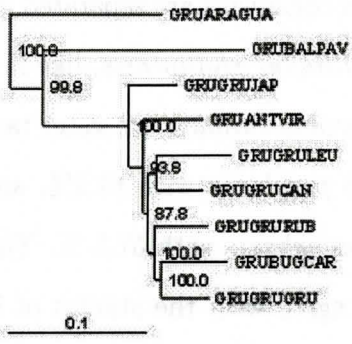

Figure 4.4: The sub-tree of cranes

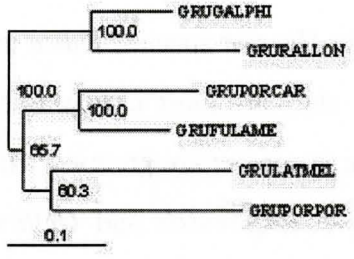

Figure 4.5: The sub-tree of rails

4.4.2 The relationships within orders and families

For the four clusters that contain non-Neoaves species, the phylogenetic trees within Tinamous, Anseriformes and Galliformes are the same in the two figures. The relationships among Tinamous is that *Eudromia elegans* and *Tinamus major* are sister clades, and then they are grouped with *Nothoprocta ornata*. The tree of Anseriformes is (*Anseranas semipalmatus*, ((*Cygnus columbianus*, *Branta canadensis*), (*Aythya americana*, *Anas platyrhynchos*))). The tree of Galliformes is shown in Figure 4.6. Regardless of the megapode, the basal position of *Ortalis guttata*, which belongs to Cracidae (guans, curassows and allies), supports the hypothesis that Cracidae is the sister clade of the group of Numididae (guineafowl), Odontophoridae (New World quail) and Phasianidae (junglefowl, pheasants, quail, and allies) (Dimcheff *et al*, 2002; Dyke *et al*, 2003).

However, the trees of Ratites by the two methods are slightly different. Figure 4.7 is the tree of Ratites created by the second method. Like Figure 4.7, the first method allied the two kiwis as a group, two moas as a group, two rheas as a group, the emu and cassowary as a group. The little difference is that the branch of Dinornithidae (moas) and Rheidae (rheas) is linked to Apterygidae (kiwis), and then the new

clade is grouped with (*Casuarius casuarius* + *Struthio camelus*) The relationships among Apterygidae, Dinornithidae and Rheidae of the second method agree with most studies of molecular data (Haddrath and Baker, 2001, Cooper *et al*, 2001)

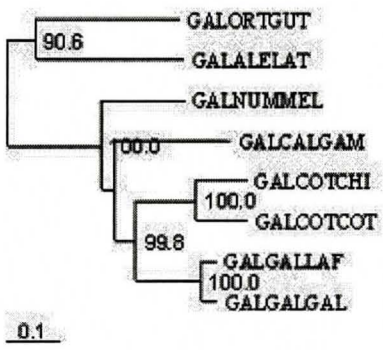

Figure 4.6: The subtree of Galliformes

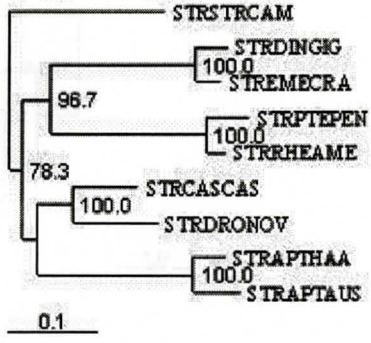

Figure 4.7: The subtree of Passeriformes

Although Passeriformes (perching birds) contains more than half of the extant birds, its relationships remain controversial except the basal clade of Acanthisittidae (new world wren) (Cracraft *et al*, 2004). However, the basal clade, Acanthisittidae, is not found in the both trees. Figure 4.8 is the phylogenetic tree of Passeriformes using the second method. Besides Passeriformes, the two methods generated the same sub-trees for most of clusters within Neoaves except Bayesian posterior probabilities.

For the both methods, the sub-tree of Falcons is ((*Falco peregrinus*, *Falco sparverius*), *Micrastur gilvicollis*), which is consistent with the both trees in the study of Griffiths *et al* (2004). The relationship of Piciformes (woodpeckers) of the two methods is ((*Picoides pubescens* + *Dryocopus pileatus*) + *pteroglossus azara*) and the Bayesian posterior probability for each branch is 100%. Within Cuculiformes (cuckoos, turacos, hoatzin), the ani, *Crotophaga ani*, is the sister clade of (*Coccyzus americanus* + *Cuculus pallidus*), which is consistent with the study of Hughes (1996)

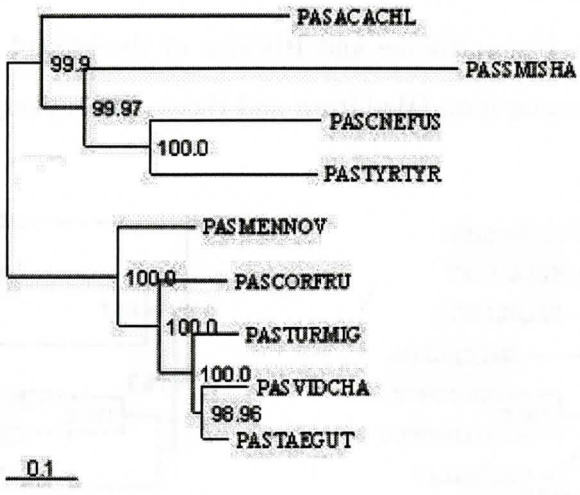

Figure 4.8: The subtree of Passeriformes

The both methods show the sister relationship between the hawk eagle, *Spizaetus alboniger*, and the buzzard, *Buteo buteo*, the sister relationship between their new clade and the vulture, *Neophron percnopterus*, and the sister relationship between the new clade of the three species and the osprey, *Pandion haliaetus*. This agrees with the sister relationship between Accipitridae (hawks and allies) and Pandionidae (osprey) (Sibley and Ahlquist, 1990; Fain and Houde 2004; Cracraft *et al*, 2004; Lerner and Mindell, 2005; Ericson *et al* 2006; Mayr, 2007) although Livezey and Zusi (2007) proposed that Falconidae (falcons and caracaras) and Pandionidae are sister clades and then they allied with Accipitridae. For the species within Bustards, the two methods show that *Afrotis afra* is the sister clade of *Ardeotis kori*, and then they are linked with *Tetrax tetrax*. The relationships within Condors and allies are also shown to be consistent by the two methods, which is the sister clade of California condor, *Gymnogyps californianus*, and Andean condor, *Vultur gryphus*, is linked with the turkey vulture, *Cathartes aura*.

For the species of Spoonbills, the two methods linked *Ajaja ajaia* and *Nipponia Nippon*, and then they are grouped with *Platalea minor*, which agree with the small sub-tree of spoonbills and ibises (Livezey and Zusi, 2007). Within Pelicans, the two species of pelicans are grouped together with 100%, and then they are linked with the shoebill stork, finally, the new clade is allied to the hamerkop heron with 100% in the two figures, which agree with some studies (van Tuinen *et al*, 2001; Cracraft *et al*, 2004). However, two different relationships among the four species of Ardeiformes are obtained using the two methods. In the sub-tree of the first methods, *Botaurus lentiginosus* is the sister clade of *Nycticorax nycticorax*, and then they are linked with *Bubulcus ibis*, finally, the group is allied to *Ardea novaehollandiae*; whereas, in the sub-tree of the second method, *Bubulcus ibis* is the sister clade of *Nycticorax nycticorax*, and then they are linked with *Ardea novaehollandiae*, finally, the group is allied to *Botaurus lentiginosus*. Except *Ardea novaehollandiae*, the sub-tree of the rest three species generated by the first method is congruent with the phylogenetic sub-tree of osteological estimate using 30 skeletal characters (McCracken and Sheldon, 1998), but the sub-tree of the three species generated by the second method is consistent with the tree of DNA-DNA hybridization and vocal estimates, the tree of osteological estimate using 33 skeletal characters (McCracken and Sheldon, 1998) and phylogenetic trees in other studies (Sheldon, 1987; Sheldon *et al* 1995). Within Gaviiformes (loons), the sub-tree of the two methods is (*Gavia stellata* + (*Gavia immer* + *Gavia pacifica*)).

Within Podicipediformes (grebes), the relationships of two methods are different. In the sub-tree of the first method, the great crested grebe, *Podiceps cristatus* is grouped with the pied-billed grebe, *Podilymbus podiceps*, and then they are linked with the western grebe, *Aechmophorus occidentalis*. Whilst in the sub-tree of the second method, the western grebe is the sister clade of the great crested grebe, then they

are grouped with the pie-billed grebe. The relationships of the second method agree with the phylogenetic study on mitochondrial sequences (Cracraft *et al*, 2004: Figure27.6), whereas the evidence for the relationships of the first method is not found. Hence, at this point, it is believed that the method using the secondary structure of rRNAs can generate more reasonable phylogenetic trees than the method without using the structure. The two methods also show the same sub-tree for Spheniscidae (penguins), which is (*Pygoscelis adeliae* + (*Eudyptula minor* + *Eudyptes chrysocome*)). For the species within Storks, the sub-tree is (*Mycteria americana*, (*Ciconia nigra*, (*Ciconia ciconia*, *Ciconia maguara*))), which agrees with the study based on the mitochondrial Cytochrome b Sequences and nuclear DNA-DNA hybridization distances (Slikas, 1997). The two methods also show that ((*Syrrhaptes paradoxus* + *Pterocles namaqua*) + *Pterocles bicinctus*) for Sandgrouse. The sub-tree of Procellariiformes (shearwaters and allies) generated by the two methods is ((*Puffinus gravis* + *Pterodroma brevirostris*) + *Diomedea melanophris*), which supports other studies (Nunn and Stanley, 1998; Kennedy and Page, 2002). The sub-tree of Apodiformes of the two methods is ((*Apus affinus*, *Apus apus*), *Anthracothorax nigricollis*), which is supported by monophyletic clades of Apodidae (swifts) and Trochilidae (hummingbirds) (Mayr, 2003; Livezey and Zusi, 2007).

Two nearly congruent phylogenetic trees of Charadriiformes were generated by the two methods except some Bayesian posterior probabilities and the relationships among the painted snipe, the seed snipe, the plainswanderer and the clade of jacana and Africana jacana. By the first method, the sub-tree of the five species is (*Attagis gayi*, (*Jacana spinosa*, *Actophilornis africanus*), (*Pedionomus torquatus*, Rostratula *bengalensis*)). The sub-tree of the second method is shown in Figure 4.9. Although the relationships among the two species of Haematopodidae (oystercatchers) and the species of Recurvirostridae (avocets and stilts) are not consistent with other stud-

ies, the sister relationship between the plover, *Charadrius vociferus*, and the lapwing, *Vanellus resplendens*, and the relationship between Charadriidae (plovers and lapwings), the golden plover, *Pluvialis dominica* and (Haematopodidae + Recurvirostridae) agree with the study of Fain and Houde (2007). Besides, the relationships among four species within Scolopacidae, *Phalaropus tricolour*, *Actitis macularius*, *Gallinago gallinago* and *Limosa fedoa*, also support the phylogenetic tree obtained in their study (Fein and Houde, 2007). Furthermore, for the both methods, the two species of buttonquails are grouped together with 100% and their new clade is close to the group of the courser, the skimmer, the murrelet, the skua/jaeger, gulls, terns, and guillemots, which agree with other studies (Paton *et al*, 2003; Fain and Houde, 2007). Although the relationships among the three species of Alcidae are not shown with high probabilities, the sub-tree among the clade, Alcidae, the courser, the two gulls, the two terns, the skimmer and the skua/jaeger is also supported by the study of Fain and Houde (2007). Moreover, the sister relationship between the painted snipe, *Rostratula bengalensis*, and the clade of jacana and Africana jacana, which is shown by the second method, agree with other studies (Livezey and Zusi, 2007; Fain and Houde, 2007). And the sub-tree of (*Rostratula bengalensis*, (*Jacana spinosa*, *Actophilornis africanus*)), the seed snipe and the plainswanderer of the second method provides another evidence for the study of Fain and Houde (2007).

4.5 General Conclusion

Unfortunately, it appears that the rRNA data is still insufficient to resolve many outstanding questions. Most of the deep level relationships within the Neoaves are not well resolved and are still controversial. In general, these results are much less informative than those obtained with the mammals using similar methods (Jow *et*

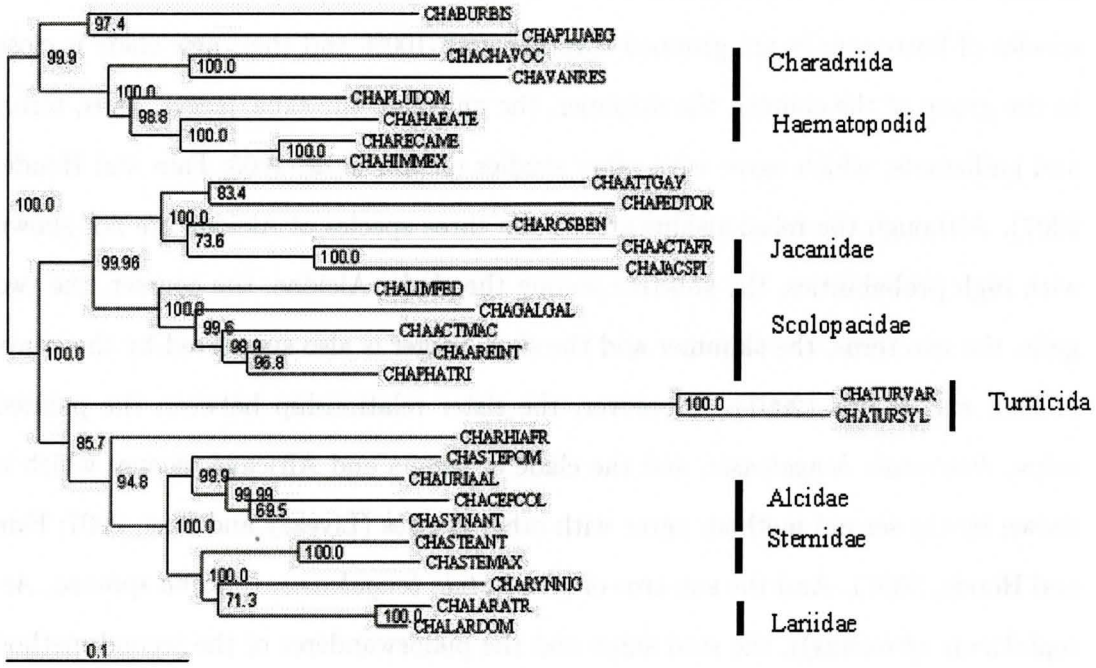

Figure 4.9: The sub-tree of Charadriiformes

al, 2002; Hudelot *et al*, 2003). Nevertheless, these results are sufficient to show that there are many problems with the traditional classification of birds.

The consensus tree of an MCMC run does not show all the alternative trees that occur within the run. Groups may therefore exist in some of these trees that are not present in the consensus tree. In this study, four orders are found to be polyphyletic in the consensus tree. We wish to test whether there is any support for these groups of species at all within the alternative trees generated by MCMC. The 'consense' program in Phylip calculates the support probability of any specified group of species within a set of alternative trees. This program was used to investigate the four polyphyletic orders.

Gruiformes: Although cranes, rails, the sungrebe and the trumpeter form a clade with probability 99.99%, there are three species, the kagu, the sunbittern and the Mesite, that have no probability to group with this clade, and the bustards group is also separate from all of these. Hence, there is strong evidence that Gruiformes are polyphyletic.

Pelecaniformes: In the initial runs before specifying the clusters, it was found that Pelecanus occidentalis and P. conspicillatus (members of Pelecaniformes) group with 100% support with the shoebill stork and hamerkop heron (members of Ciconiiformes). For these reason, these four species were treated as a cluster in the MCMC runs for which the results are presented above. This means that both Pelecaniformes and Ciconiiformes, as traditionally defined, are not monophyletic. The other groups of Pelecaniformes are the Tropic birds, Frigate birds and Cormorants. In the set of alternative trees there was 0% probability of all these groups forming a monophyletic clade with the pelicans group. Hence, there is strong evidence that Pelecaniformes is polyphyletic.

Falconiformes: The traditional classification of Falconiformes includes the clusters we have named Falcons, Hawks and allies, and Condors and allies. These three groups are all separate on the consensus tree. However there is weak support in the alternative trees for grouping Falcons with Hawks (1.5%). There is no support for grouping Condors with these two groups. In the current NCBI Taxonomy, Condors have been reclassified with Ciconiiformes. However, according to our results, the most likely relatives of Condors are Charadriiformes and bustards. Thus the classification of Condors remains extremely unclear, and whatever the position of the Condors, it seems very unlikely that the traditional order Falconiformes is monophyletic.

Ciconiiformes: We have already dealt with the shoebill stork and Hamerkop heron, which appear to be closer to pelicans than any of the other birds classed as Ciconiiformes. We have also dealt with Condors, which are classed as Ciconiiformes by NCBI and in Table 4.1. Excluding, these species, the main groups of Ciconiiformes are Storks, Spoonbills, Sandgrouse, and Ardeiformes (herons). There is 0% probability of all these groups clustering together in the alternative trees. Thus Ciconiiformes also appears to be a poorly defined, polyphyletic group.

As always with phylogenetic studies, the conclusions drawn from the analysis depend crucially on the model of molecular evolution used to calculate the likelihood of alternative trees. The method that uses the specific model for the paired sites in the RNA structure is more likely to be meaningful than the simpler model that treats all sites as single sites because it is well-established that compensatory mutations are important in RNA helices. However, there are other factors that are excluded from the model that may also be very important. One such factor is the variation of base frequencies among species, which is not accounted for in PHASE or in any of the other commonly used phylogeny programs. It is therefore clear that much larger sets

of sequences will be required, including both nuclear and mitochondrial genes, before a reliable phylogeny for birds can be obtained.

Table 4.1: 159 avian species

Genus_species	Common name	Order	Family	Abbreviation	Cluster name
Eudromia_elegans	tinamou	Tinamiformes	Tinamidae	TINEUDELE	Tinamous
Tinamus_major	tinamou	Tinamiformes	Tinamidae	TINTINMAJ	Tinamous
Nothoprocta_ornata	tinamou	Tinamiformes	Tinamidae	TINNOTORN	Tinamous
Apteryx_australis	kiwi	Apterygiformes	Apterygidae	STRAPTAUS	Ratites
Apteryx_haastii	kiwi	Apterygiformes	Apterygidae	STRAPTHAA	Ratites
Dromaius_novaehollandiae	emu	Casuariiformes	Dromaiidae	STRDRONOV	Ratites
Casuarius_casuarius	cassowary	Casuariiformes	Casuariidae	STRCASCAS	Ratites
Struthio_camelus	ostrich	Struthioniformes	Struthionidae	STRSTRCAM	Ratites
Rhea_americana	rhea	Rheiformes	Rheidae	STRRHEAME	Ratites
Pterocnemia_pennata	rhea	Rheiformes	Rheidae	STRPTEPEN	Ratites
Emeus_crassus	moa	Dinornithiformes	Emeidae	STREMECRA	Ratites
Dinornis_giganteus	moa	Dinornithiformes	Dinornithidae	STRDINGIG	Ratites
Anas_platyrhynchos	mallard	Anseriformes	Anatidae	ANSANAPLA	Anseriformes
Aythya_americana	duck	Anseriformes	Anatidae	ANSAYTAME	Anseriformes
Branta_canadensis	goose	Anseriformes	Anatidae	ANSBRACAN	Anseriformes

Cygnus_columbianus	swan	Anseriformes	Anatidae	ANSCYGCOL	Anseriformes
Anseranas_semipalmatus	magpie goose	Anseriformes	Anseranatidae	ANSANSSEM	Anseriformes
Alectura_lathami	megapode	Galliformes	Megapodiidae	GALALELAT	Galliformes
Ortalis_guttata	chacalaca	Galliformes	Cracidae	GALORTGUT	Galliformes
Gallus_gallus_gallus	chicken	Galliformes	Phasianidae	GALGALGAL	Galliformes
Gallus_lafayettei	junglefowl	Galliformes	Phasianidae	GALGALLAF	Galliformes
Coturnix_coturnix	quail	Galliformes	Phasianidae	GALCOTCOT	Galliformes
Coturnix_chinensis	quail	Galliformes	Phasianidae	GALCOTCHI	Galliformes
Numida_meleagris	Guineafowl	Galliformes	Numididae	GALNUMMEL	Galliformes
Callipepla_gambelli	american quail	Galliformes	Odontophoridae	GALCALGAM	Galliformes
Columba_livia	Pigeon	Columbiformes	Columbidae	COLCOLLIV	Single
Tyrannus_tyrannus	new world sub-oscine songbird	Passeriformes	Tyrannidae	PASTYRTYR	Passeriformes
Cnemotriccus_fuscatus	new world sub-oscine songbird	Passeriformes	Tyrannidae	PASCNEFUS	Passeriformes
Smithornis_sharpei	new world sub-oscine songbird	Passeriformes	Eurylaimidae	PASSMISHA	Passeriformes

Acanthisitta_chloris	New Zealand wren	Passeriformes	Acanthisittidae	PASACACHL	Passeriformes
Taeniopygia_guttata_B	Zebrafinch	Passeriformes	Estrildidae	PASTAEGUT	Passeriformes
Vidua_chalybaeta	oscine songbird	Passeriformes	Estrildidae	PASVIDCHA	Passeriformes
Turdus_migratorius	Robin	Passeriformes	Turdidae	PASTURMIG	Passeriformes
Corvus_frugilegus	Crow	Passeriformes	Corvidae	PASCORFRU	Passeriformes
Menura_novaehollandiae	oscine songbird	Passeriformes	Menuridae	PASMENNOV	Passeriformes
Falco_peregrinus	Falcon	Falconiformes	Falconidae	FALFALPER	Falcons
Falco_sparverius	Falcon	Falconiformes	Falconidae	FALFALSPA	Falcons
Micrastur_gilvicollis	forest falcon	Falconiformes	Falconidae	FALMICGIL	Falcons
Mesitornis_unicolor	Mesite	Gruiformes	Mesitornithidae	GRUMESUNI	Single
Phaethon_aethereus	Tropicbird	Pelecaniformes	Phaethontidae	PELPHAAET	Tropicbirds
Phaethon_rubricauda	Tropicbird	Pelecaniformes	Phaethontidae	PELPHARUB	Tropicbirds
Trogon_collaris	Trogon	Trogoniformes	Trogonidae	TROTROCOL	Single
Fregata_magnificens	Frigatebird	Pelecaniformes	Fregatidae	PELFREMAG	Frigatebirds
Fregata_minor	Frigatebird	Pelecaniformes	Fregatidae	PELFREMIN	Frigatebirds
Sula_nebouxi	Booby	Pelecaniformes	Sulidae	PELSULNEB	Cormorants and allies

Sula_leucogaster	Booby	Pelecaniformes	Sulidae	PELSULLEU	Cormorants and allies
Phalacrocorax_brasilianus	Cormorant	Pelecaniformes	Phalacrocoracidae	PELPHABRA	Cormorants and allies
Anhinga_anhinga	Snakebird	Pelecaniformes	Anhingidae	PELANHANH	Cormorants and allies
Galbula_pastazae	Jacamar	Galbuliformes	Galbulidae	GABGALPAS	Single
Tockus_nasutus	Hornbill	Bucerotiformes	Bucerotidae	BUCTOCNAS	Single
Upupa_epops	hoopoe	Upupiformes	Upupidae	UPUUPUEPO	Single
Coracias_caudata	roller	Coraciiformes	Coraciidae	CORCORCAU	Single
Picoides_pubescens	woodpecker	Piciformes	Picidae	PICPICPUB	Piciformes
Dryocopus_pileatus	woodpecker	Piciformes	Picidae	PICDRYPIL	Piciformes
Pteroglossus_azara	aracari toucan	Piciformes	Ramphastidae	PICPTEAZA	Piciformes
Coccyzus_americanus	cuckoo	Cuculiformes	Coccyzidae	CUCCOCAME	Cuculiformes
Cuculus_pallidus	cuckoo	Cuculiformes	Cuculidae	CUCCUCPAL	Cuculiformes
Crotophaga_ani	ani	Cuculiformes	Crotophagidae	CUCCROANI	Cuculiformes

Spizaetus_alboniger	hawk-eagle	Falconiformes	Accipitridae	FALSPIALB	Hawks and allies
Buteo_buteo	buzzard	Falconiformes	Accipitridae	FALBUTBUT	Hawks and allies
Neophron_percnopterus	vulture	Falconiformes	Accipitridae	FALNEOPER	Hawks and allies
Pandion_haliaetus	osprey	Falconiformes	Accipitridae	FALPANHAL	Hawks and allies
Platalea_minor	spoonbill	Ciconiiformes	Threskiornithidae	CICPLAMIN	Spoonbills
Ajaia_ajaia	ibis	Ciconiiformes	Threskiornithidae	CICAJAAJA	Spoonbills
Nipponia_nippon	ibis	Ciconiiformes	Threskiornithidae	CICNIPNIP	Spoonbills
Balaeniceps_rex	shoebill stork	Ciconiiformes	Balaenicipitidae	PELBALREX	Pelicans
Scopus_umbretta	hamerkop heron	Ciconiiformes	Scopidae	PELSCOUMB	Pelicans
Pelecanus_occidentalis	pelican	Pelecaniformes	Pelecanidae	PELPELOCC	Pelicans
Pelecanus_conspicillatus	pelican	Pelecaniformes	Pelecanidae	PELPELCON	Pelicans
Ardea_novaehollandiae	heron	Cicioniiformes	Ardeidae	ARDARDNOV	Ardeiformes
Bubulcus_ibis	heron	Cicioniiformes	Ardeidae	ARDBUBIBI	Ardeiformes
Nycticorax_nycticorax	heron	Cicioniiformes	Ardeidae	ARDNYCNYC	Ardeiformes
Botaurus_lentiginosus	heron	cicioniiformes	Ardeidae	ARDBOTLEN	Ardeiformes
Heliornis_fulica	sungrebe	Gruiformes	Heliornithidae	GRUHELFUL	Single

Fulica_americana	coot	Gruiformes	Rallidae	GRUFULAME	Rails
Porzana_carolina	rail	Gruiformes	Rallidae	GRUPORCAR	Rails
Porphyrio_porphyrio	rail	Gruiformes	Rallidae	GRUPORPOR	Rails
Laterallus_melanophaius	rail	Gruiformes	Rallidae	GRULATMEL	Rails
Rallus_longirostris	rail	Gruiformes	Rallidae	GRURALLON	Rails
Gallirallus_phillipensis	rail	Gruiformes	Rallidae	GRUGALPHI	Rails
Psophia_viridis	trumpeter	Gruiformes	Psophiidae	GRUPSOVIR	Single
Grus_grus	crane	Gruiformes	Gruidae	GRUGRUGRU	Cranes
Grus_japonensis	crane	Gruiformes	Gruidae	GRUGRUJAP	Cranes
Grus_rubicunda	crane	Gruiformes	Gruidae	GRUGRURUB	Cranes
Grus_canadensis	crane	Gruiformes	Gruidae	GRUGRUCAN	Cranes
Anthropoides_virgo	crane	Gruiformes	Gruidae	GRUANTVIR	Cranes
Bugeranus_carunculatus	crane	Gruiformes	Gruidae	GRUBUGCAR	Cranes
Grus_leucogeranus	crane	Gruiformes	Gruidae	GRUGRULEU	Cranes
Balearica_pavonina	crowned crane	Gruiformes	Gruidae	GRUBALPAV	Cranes
Aramus_guarauna	limpkin	Gruiformes	Aramidae	GRUARAGUA	Cranes
Melopsittacus_undulatus	parakeet	Psittaciformes	Psittacidae	PSIMELUND	Psittaciformes

Strigops_habroptilus	parrot	Psittaciformes	Psittacidae	PSISTRHAB	Psittaciformes
Cariama_cristata	seriema	Gruiformes	Cariamidae	GRUCARCRI	Single
Gavia_stellata	loon	Gaviiformes	Gaviidae	GAVGAVSTE	Gaviiformes
Gavia_immer	loon	Gaviiformes	Gaviidae	GAVGAVIMM	Gaviiformes
Gavia_pacifica	loon	Gaviiformes	Gaviidae	GAVGAVPAC	Gaviiformes
Eurypyga_helias	sunbittern	Gruiformes	Eurypygidae	GRUEURHEL	Single
Rhynochetos_jubata	kagu	Gruiformes	Rhynochetidae	GRURHYJUB	Single
Chordeiles_acutipennis	nighthawk	Caprimulgiformes	Caprimulgidae	CAPCHOACU	Single
Bubo_virginianus	owl	Strigiformes	Strigidae	STGBUBVIR	Strigiformes
Ninox_novaeseelandiae	owl	Strigiformes	Strigidae	STGNINNOV	Strigiformes
Phoenicopterus_ruber	flamingo	Phoenicopteriformes	Phoenicopteridae	PHOPHORUB	Single
Podiceps_cristatus	grebe	Podicipediformes	Podicipedidae	PODPODCRI	Podicipediformes
Aechmophorus_occidentalis	grebe	Podicipediformes	Podicipedidae	PODAECOCC	Podicipediformes
Podilymbus_podiceps	grebe	Podicipediformes	Podicipedidae	PODPODPOD	Podicipediformes
Eudyptes_chrysocome	penguin	Sphenisciformes	Spheniscidae	SPHEUDCHR	Sphenisciformes
Eudyptula_minor	penguin	Sphenisciformes	Spheniscidae	SPHEUDMIN	Sphenisciformes
Pygoscelis_adeliae	penguin	Sphenisciformes	Spheniscidae	SPHPYGADE	Sphenisciformes

Ciconia_ciconia	stork	Ciconiiformes	Ciconiidae	CICCICCIC	Storks
Ciconia_maguara	stork	Ciconiiformes	Ciconiidae	CICCICMAG	Storks
Mycteria_americana	stork	Ciconiiformes	Ciconiidae	CICMYCAME	Storks
Ciconia_nigra	stork	Ciconiiformes	Ciconiidae	CICCICNIG	Storks
Syrrhaptes_paradoxus	sandgrouse	Ciconiiformes	Pteroclidae	COLSYRPAR	Sandgrouse
Pterocles_namaqua	sandgrouse	Ciconiiformes	Pteroclidae	COLPTENAM	Sandgrouse
Pterocles_bicinctus	sandgrouse	Ciconiiformes	Pteroclidae	COLPTEBIC	Sandgrouse
Puffinus_gravis	shearwater	Procellariiformes	Procellariidae	PROPUFGRA	Procellariiformes
Pterodroma_brevirostris	petrel	Procellariiformes	Procellariidae	PROPTEBRE	Procellariiformes
Diomedea_melanophris	albatross	Procellariiformes	Diomedeidae	PRODIOMEL	Procellariiformes
Apus_affinus	swift	Apodiformes	Apodidae	APOAPUAFF	Apodiformes
Apus_apus	swift	Apodiformes	Apodidae	APOAPUAPU	Apodiformes
Anthracothorax_nigricollis	hummingbird	Trochiliformes	Trochilidae	APOANTNIC	Apodiformes
Musophaga_violaceae	turaco	Musophagiformes	Musophagidae	MUSMUSVIO	Single
Opisthocomus_hoazin	hoatzin	Opisthocomiformes	Opisthocomidae	HOAOPIHOA	Single
Afrotis_afra	bustard	Gruiformes	Otididae	RUAFRAFR	Bustards
Ardeotis_kori	bustard	Gruiformes	Otididae	GRUARDKOR	Bustards

Tetrax_tetrax	bustard	Gruiformes	Otididae	GRUTETTET	Bustards
Gymnogyps_californianus	california condor	Ciconiiformes	Cathartidae	FALGYMCAL	Condors and allies
Vultur_gryphus	andean condor	Ciconiiformes	Cathartidae	FALVULGRY	Condors and allies
Cathartes_aura	turkey vulture	Ciconiiformes	Cathartidae	FALCATAUR	Condors and allies
Burhinus_bistriatus	thickknee	Charadriiformes	Burhinidae	CHABURBIS	Charadriiformes
Charadrius_vociferus	plover	Charadriiformes	Charadriidae	CHACHAVOC	Charadriiformes
Vanellus_resplendens	lapwing	Charadriiformes	Charadriidae	CHAVANRES	Charadriiformes
Haematopus_ater	oystercatcher	Charadriiformes	Haematopodidae	CHAHAEATE	Charadriiformes
Himantopus_mexicanus	oystercatcher	Charadriiformes	Haematopodidae	CHAHIMMEX	Charadriiformes
Recurvirostra_americana	avocet	Charadriiformes	Recurvirostridae	CHARECAME	Charadriiformes
Pluvialis_dominica	golden plover	Charadriiformes	Charadriidae	CHAPLUDOM	Charadriiformes
Pluvianus_aegyptius	egyptian plover	Charadriiformes	Glareolidae	CHAPLUAEG	Charadriiformes
Phalaropus_tricolor	phalarope	Charadriiformes	Scolopacidae	CHAPHATRI	Charadriiformes
Arenaria_interpres	turnstone	Charadriiformes	Scolopacidae	CHAAREINT	Charadriiformes

Actitis_macularius	sandpiper	Charadriiformes	Scolopacidae	CHAACTMAC	Charadriiformes
Gallinago_gallinago	snipe	Charadriiformes	Scolopacidae	CHAGALGAL	Charadriiformes
Limosa_fedoa	godwit	Charadriiformes	Scolopacidae	CHALIMFED	Charadriiformes
Jacana_spinosa	jacana	Charadriiformes	Jacanidae	CHAJACSPI	Charadriiformes
Actophilornis_africanus	africana jacana	Charadriiformes	Jacanidae	CHAACTAFR	Charadriiformes
Rostratula_bengalensis	painted snipe	Charadriiformes	Rostratulidae	CHAROSBEN	Charadriiformes
Pedionomus_torquatus	plainswanderer	Charadriiformes	Pedionomidae	CHAPEDTOR	Charadriiformes
Attagis_gayi	seedsnipe	Charadriiformes	Thinocoridae	CHAATTGAY	Charadriiformes
Turnix_sylvatica	buttonquail	Turniciformes	Turnicidae	CHATURSYL	Charadriiformes
Turnix_varia	buttonquail	Turniciformes	Turnicidae	CHATURVAR	Charadriiformes
Rhinoptilus_africanus	courser	Charadriiformes	Glareolidae	CHARHIAFR	Charadriiformes
Larus_dominicanus	gull	Charadriiformes	Lariidae	CHALARDOM	Charadriiformes
Larus_atricilla	gull	Charadriiformes	Lariidae	CHALARATR	Charadriiformes
Sterna_maxima	tern	Charadriiformes	Lariidae	CHASTEMAX	Charadriiformes
Sterna_antillarum	tern	Charadriiformes	Lariidae	CHASTEANT	Charadriiformes
Rynchops_niger	skimmer	Charadriiformes	Lariidae	CHARYNNIG	Charadriiformes
Syntliboramphus_antiquus	murrelet	Charadriiformes	Alcidae	CHASYNANT	Charadriiformes

Uria_aalge	guillemot	Charadriiformes	Alcidae	CHAURIAAL	Charadriiformes
Cepphus_columba	guillemot	Charadriiformes	Alcidae	CHACEPCOL	Charadriiformes
Stercorarius_pomarinus	skua/jaeger	Charadriiformes	Stercorariidae	CHASTEPOM	Charadriiformes

Bibliography

[1] Akashi H, 2003, Translational selection and yeast proteome evolution. *Genetics* 164: 1291-1303.

[2] Akashi, H, 1994, Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. *Genetics* 136: 927-935.

[3] Alberts B, Johnson A, Lewis J, Raff M, Roberts k and Walter P, 2002, Garland Publishing, a member of the Taylor & Francis Group, *Molecular Biology of the Cell, 4th Edition*

[4] Antezana MA, Kreitman M, 1999, The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences,*J. Mol. Evol*, 49: 36-43

[5] Arndt PF, Hwa T, 2005, Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics.* 21:2322-2328.

[6] Benson DA, Boguski M, Lipman DJ and Ostell J, 1994, GenBank, *Nucleic Acids Research* 22: 3441 -3444

[7] Bielawski JP, Gold JR, 2002, Mutation patterns of mitochondrial H- and L-strand DNA in closely-related cyprinid fishes. *Genetics* 161: 1589-1597.

[8] Bogenhagen DF and Clayton DA, 1977, Mouse L cell mitochondrial DNA molecules are selected randomly for replication throughout the cell cycle, *Cell* 11, 719–727.

[9] Braun EL and Kimball RT, 2002, Examining basal avian divergences with mitochondrial sequences: model complexity, taxon sampling, and sequence length. *Systematic Biology*, 51: 614–625

[10] Brown TA, Cecconi C, Tkachuk AN, Bustamante C and Clayton DA, 2005, Replication of mitochondrial DNA occurs by strand displacement with alternative light-strand origins, not via a strand-coupled mechanism, *Genes & Dev.* 19: 2466-2476

[11] Burnham KP, Anderson DR, 1998, Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, USA.

[12] Castro JA, Picornell A and Ramon M, 1998, Mitochondrial DNA: a tool for populational genetics studies, *Internal Microbiol* 1:327–332

[13] Cavalli-Sforza L.L. and and Edwards A.W.F, 1967, Phylogenetic analysis: Models and estimation procedures,*Evol.* 21: 550-570.

[14] Chan DC, 2006, Mitochondria: Dynamic Organelles in Disease, Aging, and Development, *Cell*, 125: 1241-1252

[15] Chubb AL, 2004, New nuclear evidence for the oldest divergence among neognath birds: the phylogenetic utility of ZENK (i). *Mol Phyl Evol* 30:140–151

[16] Codd, E.F, 1970, A Relational Model of Data for Large Shared Data Banks, *Communications of the ACM*, 13: 377–387

[17] Connolly TM and Begg CE, 1999, Addison Wesley Longman Limited, *Database Systems: A Practical Approach to Design, Implementation and Management, 2nd Edition*

[18] Cooper A, Lalueza-Fox C, Anderson S, Rambaut A, Austin J, Ward R, 2001, Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution, *Nature* 409: 704–707

[19] Cracraft J, 1981, Toward a phylogenetic classification of the recent birds of the world (Class Aves). *Auk* 98: 681–714.

[20] Cracraft J and Donoghue MJ, eds, 2004, New York: Oxford University Press, *Assembling the tree of life.*

[21] dos Reis M, Wernisch L, Savva R, 2003, Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res.* 31: 6976-6985.

[22] Duret L, 2002, Evolution of synonymous codon usage in metazoans. Curr. Op. *Genet. Dev.* 12: 640-649.

[23] Dyke GJ, Gulas BE, Crowe TM, 2003, Suprageneric relationships of galliform birds (Aves, Galliformes): a cladistic analysis of morphological characters. *Zoological Journal of the Linnean Society* 137: 227–244

[24] Dyke, G. J., and van Tuinen M, 2004, The evolutionary radiation of modern birds (Neornithes) reconciling molecules morphology and the fossil record. *Zool. J. Linn. Soc.* 141:153–177.

[25] Ericson PGP, Anderson CL, Britton T, Elzanowski A, Johansson US, Ka llersjo M, Ohlson JI, Parsons TJ, Zuccon D, Mayr G, 2006, Diversification of Neoaves: integration of molecular sequence data and fossils. *Biol Lett* 2:543–547.

[26] Fain MG and Houde P,2004, Parallel radiations in the primary clades of birds. *Evolution* 58: 2558–2573

[27] Fain MG, Krajewski C and Houde P, 2007, Phylogeny of "core Gruiformes" (Aves: Grues) and resolution of the Limpkin–Sungrebe problem, *Molecular Phylogenetics and Evolution* 43:515–529

[28] Fain MG and Houde P, 2007, Multilocus perspectives on the monophyly and phylogeny of the order Charadriiformes (Aves), *BMC Evolutionary Biology* , 7:35-50

[29] Faith JJ, Pollock DD, 2003, Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* 165: 735-745.

[30] Fedorov A, Saxonov S, Gilbert W, 2002, Regularities of context-dependent codon bias in eukaryotic genes. *Nucleic Acids Res.* 30: 1192-1197

[31] Felsenstein J. 2004, Sinauer Associates, Sunderland, MA ,*Inferring Phylogenies.*

[32] Fernández-Silva P, Enriquez JA and Montoya J, 2003, Special Review Series – Biogenesis and Physiological Adaptation of Mitochondria Replication and transcription of mammalian mitochondrial DNA, *Experimental Physiology* 88: 41–56

[33] Fitch W.M. and Margoliash E, 1967, Construction of Phylogenetic Trees, Science, 155: 279 – 284

[34] Foster PG, Jermiin LS, Jickey DA, 1997, Nucleotide compositional bias affects amino acid frequencies in proteins coded by animal mitochondria. *J. Mol. Evol.* 44: 282-288.

[35] García-Moreno J, Sorenson MD, Mindell DP, 2003, Congruent avian phylogenies inferred from mitochondrial and nuclear DNA sequences, *Journal of Molecular Biology* 57: 27–37

[36] Gibson, A. Gowri-Shankar, V., Higgs, P.G. & Rattray, M., 2005, A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol. Evol. 22, 251-264.

[37] Griffiths CS, Barrowclough GF, Groth JG, Mertz L, 2004, Phylogeny of the Falconidae (Aves): a comparison of the efficacy of morphological, mitochondrial, and nuclear data, *Molecular Phylogenetics and Evolution*, 32:101–109

[38] Gowri-Shankar V and Jow H, 2006, PHASE: a Software Package for Phylogenetics And Sequence Evolution

[39] Groth JG and Barrowclough GF, 1999, Basal divergences in birds and the phylogenetic utility of the nuclear RAG-1 gene, *Molecular Phylogenetics and Evolution* 12: 115–123

[40] Haddrath O, Baker AJ, 2001, Complete mitochondrial DNA genome sequences of extinct birds: ratite phylogenetics and the vicariance biogeography hypothesis, *Proceedings of the Royal Society of London (Series B)* 268: 939–945.

[41] Harrison GL, McLenachan PA, Phillips MJ, Slack KE, Cooper A, Penny D., 2004, Four new avian mitochondrial genomes help get to basic evolutionary questions in the late Cretaceous. Molecular Biology and Evolution 21: 974–983.

[42] Higgs, P. G., 2000, Rna secondary structure: physical and computational aspects. *Quart. Rev. of Bioph.*, 22:199–253.

[43] Higgs PG, Jameson D, Jow H. Rattray M, 2003, The evolution of tRNA-Leucine genes in animal mitochondrial genomes. *J. Mol. Evol.* 57: 435-445.

[44] Higgs PG, Hao W, Golding GB, 2007, Identification of selective effects on highly expressed genes. *Evolutionary Bioinformatics.* 2: 1-13.

[45] Hudelot, C., Gowri-Shankar, V., Jow, H., Rattray, M. and Higgs, P.G., 2003, RNA-based Phylogenetic Methods: Application to Mammalian Mitochondrial RNA Sequences. *Mol. Phyl. Evol.* 28, 241-252.

[46] Hughes JM, 1996, Phylogenetic analysis of the Cuculidae (Aves, Cuculiformes) using behavioral and ecological characters, *Auk* 113: 10–22

[47] Huxley TH, 1867, On the classification of birds; and on the taxonomic value of the modifications of certain of the cranial bones observable in that class. *Proceedings of the Zoological Society of London* 1867: 415–472.

[48] Ikemura T, 1981, Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* 151: 389-409.

[49] Jameson D, Gibson AP, Hudelot Cand Higgs, PG, 2003, OGRe: a relational database for comparative analysis of mitochondrial genomes, *Nucleic Acids Research*, 31: 202–206

[50] Jameson D,2004, *the Comparative Analysis of Mitochondrial Genome,* A thesis of PhD degree of University of Manchester

[51] Jansen RP, 2000, Origin and persistence of the mitochondrial genome. *Hum Reprod. 15 Suppl* 2:1-10.

[52] Jermiin LS, Graur D, Crozier RH, 1995, Evidence from analyses of intergenic regions for strand-specific directional mutation pressure in metazoan mitochondrial DNA. *Mol Biol Evol*, 12(4):558-563.

[53] Johansson US, Ericson PGP, 2003, Molecular support for a sister group relationship between Pici and Galbulae (Piciformes sensu Wetmore 1960). *Journal of Avian Biology* 34: 185–197.

[54] Jow H, Hudelot C,Rattray M and Higgs P, 2002, Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution, *Molecular Biology and Evolution*, 19:1591-1601

[55] Karlin S, Mrazek J, 1996, What drives codon choices in human genes? *J. Mol. Biol.* 262: 459-472.

[56] Karlin S, Mrazek J, 1997, Compositional differences within and between eukaryotic genomes. *Proc. Acad. Nat. Sci.* USA 94: 10227-10232.

[57] Kennedy M and Page RDM, 2002, Seabird supertrees: combining partial estimates of procellariform phylogeny. *Auk* 119: 88–108

[58] Knight RJ, Freeland SJ, Landweber LF, 2001, Rewiring the keyboard: evolvability of the genetic code. *Nature Rev Genet* 2:49–58

[59] Krajewski C and Dickerman AW, 1990, Bootstrap analysis of phylogenetic trees derived from DNA hybridization distances. *Syst. Zool.* 39:383-390

[60] Krajewski C, and Fetzner Jr, 1994, Phylogeny of cranes (Gruiformes: Gruidae) based on cytochrome-b DNA sequences. *Auk* 111, 351–365.

[61] Larry W, Christiansen T and and Orwant J, 2000, O'Reilly, *Programming Perl, 3rd Edition*

[62] Lerner HRL and Mindell DP, 2005, Phylogeny of eagles, Old World vultures, and other Accipitridae based on nuclear and mitochondrial DNA, *Molecular Phylogenetics and Evolution* 37: 327–346

[63] Livezey BC, 1998, A phylogenetic analysis of the Gruiformes (Aves) based on morphological characters, with an emphasis on the rails (Rallidae). *Philos Trans R Soc Lond B Biol Sci* 353:2077–2151.

[64] Livezey BC and Zusi RL, 2001, Higher-order phylogenetics of modern Aves based on comparative anatomy, *Netherlands Journal of Zoology* 51: 179–206

[65] Livezey BC, Zusi RL, 2007, Higher-order phylogeny of modern birds (Theropoda, Aves: Neornithes) based on comparative anatomy. II. Analysis and discussion. *Zool J Linn Soc* 149:1–95.

[66] Manegold A, 2006, Two additional synapomorphies of grebes Podicipedidae and flamingos Phoenicopteridae. *Acta Ornithol* 41:79–82.

[67] Maria Falkenberg, Nils-Goran Larsson, and Claes M. Gustafsson, 2007, Transcription in Mammalian Mitochondria, *Annu. Rev. Biochem.*76:679–99.

[68] Mayr G and Clarke J, 2003, The deep divergences of neornithine birds: a phylogenetic analysis of morphological characters. *Cladistics* 19: 527–553.

[69] Mayr G, 2003, The phylogenetic relationships of the shoebill, Balaeniceps rex. *J Ornithol* 144:157–175.

[70] Mayr G, 2004, Morphological evidence for sister group relationship between flamingos (Aves: Phoenicopteridae) and grebes (Podicipedidae), *Zool J Linn Soc* 140:157–169.

[71] Mayr G, 2007, Avian higher-level phylogeny: well-supported clades and what we can learn from a phylogenetic analysis of 2954 morphological characters, *J Zool Syst Evol Res* doi: 10.1111/j.1439-0469.

[72] McCracken KG and Sheldon FH, 1998, Molecular and Osteological Heron Phylogenies: Sources of Incongruence, *The Auk*, 115:127-141.

[73] Morton BR, Wright SI, 2007, Selective constraints on codon usage of nuclear genes from Arabidopsis thaliana. *Mol. Biol. Evol.* 24: 122-129.

[74] Nijssen GM and Halpin TA, 1989, Prentice-Hall, Inc. *Conceptual schema and relational database design: a fact oriented approach*

[75] Nunn GB and Stanley SE, 1998, Body size effects and rates of cytochrome b evolution in tube-nosed seabirds, *Molecular Biology and Evolution* 15: 1360–1371.

[76] Olson, S. L., 1983, Evidence for a polyphyletic origin of the Piciformes. *Auk* 100, 126–133.

[77] Osawa S, 1995, Oxford University Press, Oxford, *Evolution of the Genetic Code*

[78] Paton TA, Baker AJ, Groth JG, Barrowclough GF, 2003, RAG-1 sequences resolve phylogenetic relationships within charadriiform birds. *Mol Phyl Evol* 29:268–278.

[79] Raikow RJ and Cracraft J, 1983, Monophyly of the Piciformes: A Reply to Olson, *The Auk*, 100: 134-138.

[80] Reyes A, Gissi C, Pesole G, Saccone C, 1998, Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol. Biol. Evol.* 15: 957-966.

[81] Sengupta S, Yang X, Higgs PG, 2007, The mechanisms of codon reassignments in mitochondrial genetic codes. (in press *J. Mol. Evol.*)

[82] Shadel, G.S. and Clayton, D.A, 1997, Mitochondrial DNA maintenance in vertebrates. *Annu. Rev. Biochem.* 66: 409–435.

[83] Sharp PM, Li WH, 1987, The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15: 1281-1295.

[84] Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F,1988, Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within species diversity. *Nucleic Acids Res.* 16: 8207-8211.

[85] Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE, 2005, Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 33: 1141-1153.

[86] Sheldon FH, 1987, Phylogeny of herons estimated from DNA-DNA hybridization data. *Auk* 104:97-108.

[87] Sheldon FH, McCracken KG and Stuebing KD, 1995, Phylogenetic relationships of the Zigzag Heron (Zebrilus undulatus) and Whitecrested Bittern (Tigriornis leucolophus) estimated by DNA-DNA hybridization, *Auk* 112:672-679

[88] Shioiri C, Takahata N, 2001. Skew of mononucleotide frequencies, relative abundance of dinucleotides and DNA strand asymmetry. *J. Mol. Evol.* 53: 364-376.

[89] Sibley CG and Ahlquist JE, 1990, New Haven, CT: Yale University Press, *Phylogeny and classification of birds: a study in molecular evolution*

[90] Simmons MP and MiyaM, 2004, Efficiently resolving the basal clades of a phylogenetic tree using Bayesian and parsimony approaches: a case study using mitogenomic data from 100 higher teleost fishes. *Mol Phyl Evol* 31:351–362

[91] Singer GAC, Hickey DA, 2000, Nucleotide bias causes a genome-wide bias in amino acid composition of proteins. *Mol. Biol. Evol.* 17: 1581-1588.

[92] Slikas B, 1997, Phylogeny of the avian family Ciconiidae (storks) based on cytochrome b sequences and DNA-DNA hybridization distances. *Molecular Phylogenetics and Evolution* 8: 275–300.

[93] Stoletzki N, Eyre-Walker A,2007, Synonymous codon usage in *Escherichia coli*: Selection for translational accuracy. *Mol. Biol. Evol.* 24: 374-381.

[94] Tanaka M and Ozawa T,1994, Strand asymmetry in human mitochondrial DNA mutations. *Genomics* 22:327–335.

[95] Tavare S, 1986, Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect. Math. Life Sc.*, 17:262–272

[96] Urbina D, Tang B, Higgs PG.,2006, The response of amino acid frequencies to directional mutational pressure in mitochondrial genome sequences is related

to the physical properties of the amino acids and to the structure of the genetic code. *J. Mol. Evol.* 62: 340-361.

[97] van Tuinen M, Sibley CG. and Hedges SB, 1998, Phylogeny and biogeography of ratite birds inferred from DNA sequences of the mitochondrial ribosomal genes. *Mol. Biol. Evol.* 15:370–376

[98] van Tuinen M, Sibley CG, Hedges SB, 2000, The early history of modern birds inferred from DNA sequences of nuclear and mitochondrial ribosomal genes. *Molecular Biology and Evolution* 17: 451–457.

[99] van Tuinen M, Butvill DB, Kirsch JAW, Hedges SB, 2001, Convergence and divergence in the evolution of aquatic birds. *Proceedings of the Royal Society of London (Series B)* 268: 1–6

[100] Watanabe M, Nikaido M, Tsuda TT, Inoko H, Mindell DP, Murata K, Okada N, 2006, The rise and fall of the CR1 subfamily in the lineage leading to penguins, *Gene* 365: 57–66

[101] Wiesner RJ, Ruegg JC, Morano I, 1992, Counting target molecules by exponential polymerase chain reaction, copy number of mitochondrial DNA in rat tissues. *Biochim Biophys Acta.* 183: 553–559.

[102] Wolstenholme, DR, 1992, Animal mitochondrial DNA: structure and evolution. *Int. Rev. Cytol.* 141: 173-216

[103] Wright F, 1990, The effective number of codons used in a gene. *Gene* 87: 23-29.

[104] Xia XH,2005, Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. *Gene* 345: 13-20.

[105] Xu, W., Jameson, D., Tang, B., Higgs, P.G., 2006, The relationship between the rate of molecular evolution and the rate of genome rearrangement in animal mitochondrial genomes. *J.Mol. Evol*, 63: 375-392

[106] Yang Z, 1994, Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.*, 39:306–314.

[107] Zwickl DJ and Hillis DM, 2002, Increased taxon sampling greatly reduces phylogenetic error, *Systematic Biology*, 51: 588-598

4147 100