NOTE TO USERS

uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Huai-Chun Wang
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Biology)
GRADE / DEGREE

Department of Biology
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

The Effects of Nucleotide Bias on Genome Evolution

TITRE DE LA THÈSE / TITLE OF THESIS

Donal Hickey
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

George Carmody

Guy Drouin

Paul Liu

Marcel Turcotte

Xuhua Xia

Gary W. Slater
LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES /
DEAN OF THE FACULTY OF GRADUATE AND POSTDOCORAL STUDIES

The Effects of Nucleotide Bias on Genome Evolution

by

Huai-Chun Wang, B.Med., M.Med.

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
In partial fulfillment of the requirements for the Ph.D. degree
in the Ottawa-Carleton Institute of Biology

Thèse soumise à
Faculté des études supérieures et postdoctorales
Université d'Ottawa
En vue de l'obtention du doctorat ès sciences
L'Institut de Biologie d'Ottawa-Carleton

May 12 2005

Your file Votre référence
ISBN: 0-494-11033-3
Our file Notre référence
ISBN: 0-494-11033-3

To the memory of my father

Acknowledgements

I would like to thank my supervisor, Professor Donal Hickey for assigning me this interesting thesis topic and directing every detail of the whole thesis work, and lots of financial support for attending conferences and summer schools. I also thank the members of my advisory committee, Drs. George Carmody, Guy Drouin and Marcel Turcotte for their advice. Dr. Xuhua Xia also gave very insightful advice that is greatly appreciated. Ada Chyurlia always gave me help in need and Dr. Greg Singer helped a lot by maintaining the Sun server (even after he had left the lab), which was essential for my programming. Financial supports from NSERC (to Dr. Hickey), OGS and University of Ottawa graduate scholarships are appreciated. Finally, as a student at a senior age, family supports are all important. I would like to thank my parents for bringing me up and my brothers and sister for taking care of them so I can study uninterrupted for the past three and a half years. My wife Shuli and my son Sean gave me love and joy so that my study was less stressful.

Abstract

The genomic G+C content of prokaryotes varies from approximately 23% to 77% among genomes. In contrast, among vertebrates, the variation is greatest within the same genome rather than between genomes. There has been a long-standing controversy concerning the causes of these inter- and intra-specific variations. Is it caused by natural selection, favored by the selectionists or, conversely, is it selectively neutral (the neutralist view)? In this study, we investigated the source of nucleotide compositional variation (nucleotide bias) and the consequences of the bias on protein sequence and genome evolution. Thermal adaptation is a primary example to study the effect of natural selection and has been thoroughly studied in this project. We found that both GC content and length of ribosomal RNA genes show positive correlations with optimal growth temperature in prokaryotes and these correlations are not due to phylogenetic history. The correlations are concentrated almost entirely within the stem regions of the rRNA. The rRNA loops, however, show very constant base composition regardless of temperature optima or genomic GC content. The loops were found to have very high amount of adenosine nucleotides throughout prokaryotes and eukaryotes. These results clearly demonstrated that environmental temperature is a selective force that drives rRNA gene evolution and different segments of the same gene (i.e., the stems and loops of the rRNA gene) experience differential selection, although the mutation spectrum presumably should be similar between the loops and stems.

For protein coding genes, mutation and natural selection play a different role compared to the rRNA genes. The neutralist predicts mutational bias would cause protein sequence evolution, while the selectionist would predict that the protein sequence is not related to genomic GC content. To investigate these two postulations and analyze the consequences of nucleotide bias in eukaryotic genomes, we studied homologous genes and their encoded proteins in two flowering plants, *Oryza sativa* (rice) and *Arabidopsis thaliana*. While there is a relatively homogenous GC content in the *Arabidopsis* genes (26% to 69%), the GC content of the rice genes is very heterogeneous (27% to 90%). High GC rice genes encode proteins having a high frequency of GC-rich codons encoded amino acids, *i.e.*, glycine, alanine, arginine and proline. Low GC rice genes and

Arabidopsis genes encode proteins having a high frequency of AT-rich codons encoded amino acids, *i.e.*, phenylalanine, tyrosine, methionine, isoleucine, asparagines and lysine. Furthermore, the effects of nucleotide bias on synonymous codon usage in the rice and *Arabidopsis* genomes were studied. We have shown that synonymous codon usage in the rice genome is primarily dictated by the GC content of the genes, rather than by translational selection. This study in multicellular higher plants, together with previous work on prokaryote and yeast, provide persuasive evidence that mutational nucleotide bias is a cause, rather than a consequence, of protein evolution and this affects codon usage and protein composition in a predictable way.

Résumé

Le contenu génomique en nucléotides G+C des prokaryotes varie approximativement de 23% à 77% entre génomes. Chez les vertébrés, contrairement aux prokaryotes, cette variation est plus élevée au sein d'un génome plutôt qu'entre génomes différents. Il existe une controverse de longue durée concernant les causes de ces variations inter- et intra-spécifique. Est-ce que ces variations sont le résultat de la sélection naturelle, cette explication est favorisée par les sélectionnistes ou d'une sélection neutre (point vue des neutralistes)? Dans cette étude, nous examinons la source des variations de la composition en nucléotides (biais nucléotidique) et les conséquences de ce biais sur les séquences de protéines et sur l'évolution du génome. L'adaptation thermique convient bien à l'étude de l'effet de la sélection naturelle et nous l'avons étudié de façon rigoureuse dans ce projet. Nous avons trouvé que le contenu en nucléotides GC et la longueur des gènes d'ARN ribosomal (ARNr) montrent une corrélation positive avec les températures optimales de croissance chez les prokaryotes et que ces corrélations ne sont pas le résultat de l'histoire phylogénétique des espèces. Les corrélations sont concentrées principalement au niveau des tiges des ARNr. Par contre, les bras d'ARNr sont très constants dans leur composition en nucléotides et ne sont pas affectés par les températures optimales ou le contenu en nucléotides GC. Il semblerait que les bras abondent de nucléotides d'adénosines et ce autant chez les procaryotes que chez les eucaryotes. Ces résultats démontrent clairement que la température environnementale exerce une force sélective qui conduit à l'évolution des gènes ARNr et que différents segments du même gène (i.e., les tiges et les bras du gène d'ARNr) sont affecté par différente sélection même si le spectre des mutations est présumé être similaire dans les tiges et les bras.

Pour les gènes codant pour des protéines, les mutations et la sélection naturelle jouent un rôle différent comparé aux gènes ARNr. Les neutralistes prédissent qu'un biais mutationnelle chez les séquences protéiques n'est pas associé au contenu en GC génomique. L'investigation de ces deux postulats et analyser les conséquences du biais nucléotidique, nous avons étudie des gènes homologues et qui codent pour des protéines de deux plantes à fleurs: *Oryza sativa* (riz) et *Arabidopsis thaliana*. Pendant que le contenu en GC est relativement homogène dans les gènes d'*Arabidopsis* (26% à 69%), le

contenu en GC chez le riz est très hétérogène (27% à 90%). Les gènes de riz à haute teneur en GC codent pour des protéines dont la fréquence d'acides aminés ayant des codons riches en nucléotides GC est élevée, i.e. glycine, alanine, arginine et proline. Chez le riz et *Arabidopsis*, les gènes à plus faible teneur en GC codent pour des protéines composé d'acide aminés ayant des codons riches en base AT, i.e, phenylalanine, tyrosine, methionine, isoleucine, asparagines et lysine. De plus, les effets du biais nucléotidique sur l'utilisation des codons synonymes du génome du riz est principalement contrôlé par le contenu en nucléotides GC des gènes, plutôt que par la sélection traductionelle. Cette étude des plantes multicellulaires d'ordre supérieur, de même que des recherches passées sur les prokaryotes et levures, démontrent de façon claire et précise que le biais nucléotidique est une cause plutôt qu'une conséquence de l'évolution des protéines et ceci affecte l'utilisation des codons et la composition protéique que nous pouvons prédire.

List of Abbreviations

(see also IUPAC code table for amino acid and nucleotide acid abbreviations)

AT (i.e., A+T) content: molar ratio of adenine and thymine in a DNA sequence, i.e. 1-G+C%.

GC (i.e., G+C) content: molar ratio of guanine and cytosine over all nucleotides. We use G+C and GC (and also A+T and AT) content interchangeably throughout the thesis.

GC1: G+C content at first codon positions

GC2: G+C content at second codon positions

GC3: G+C content at third codon positions

GC3s: G+C content at third synonymous codon positions (excluding Met, Trp and stop codons)

GCintron: G+C content of an intron

BGC: biased gene conversion

bp: base pair

kb: kilo base

CAI: codon adaptation index

CDS: coding sequence

dN: mean nonsynonymous substitutions per non-synonymous site

dS: mean synonymous substitutions per synonymous site

EMBL: DNA database of European Molecular Biology Laboratory

EMBOSS: European Molecular Biology Open Software Suite

FTP: (internet) file transfer protocol

Meso: mesophiles (mesophilic species)

Nc: effective number of codons, also called ENc

Ncprime: a modified Nc that takes into account the background nucleotide composition

NCBI: National Center for Biotechnology Information

NS: non-significant

OPT (Topt): optimal growth temperature

PAML: Phylogenetic Analysis by Maximum Likelihood

RC: (codon) redundancy class

rRNA: ribosomal RNA

RSCU: Relative Synonymous Codon Usage

ssu rRNA: small subunit rRNA

Thermo: thermophiles (thermophilic species)

tRNA: transfer RNA

TT: thymine dimer

UV: ultraviolet

VSP: (DNA repair) very short patch

IUPAC code table

Amino acid codes

1-letter	3-letter	description
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine
Z	Glx	Glutamine or Glutamic acid

Nucleic acid codes

code	description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base (A, C, G, T, or U)

Table of Contents

Acknowledgements....iii
Abstract....iv
Résumé....vi
List of Abbreviations....viii
IUPAC Code Table....x
List of Tables....xv
List of Figures....xvii

Chapter 1 General introduction....1
1.1 Variations (bias) in GC content....1
1.2 Effects of GC variations....2
1.3 Biased DNA mutation and natural selection shape GC content....5
1.3.1 Selectionist interpretations of GC content change....5
1.3.2 Neutralist interpretations of genomic differences in GC content....7
1.3.3 DNA mutation....9
1.3.4 Bias in mutation....11
1.3.5 DNA mutation repair and repair bias....12
1.3.6 Interaction of neutral evolution and selection....13
1.4 Research proposal....14
1.5 Comparative methods....15
1.6 Organization of the thesis....18

Chapter 2 Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes....21
2.1 Abstract....21
2.2 Introduction....22
2.3 Materials and Methods....23
2.4 Results....24

2.4.1 Average nucleotide composition in mesophiles and thermophiles.....24
2.4.2 16S rRNA stems and loops are affected very differently by growth temperature.....28
2.4.3 The relationship between the nucleotide content of the 16S rRNA and the nucleotide content of the whole genome.....31
2.5 Discussion.....34

Chapter 3 Thermal adaptation of ribosomal RNA genes.....39
3.1 Abstract.....39
3.2 Introduction.....40
3.3 Methods.....42
3.3.1 Sequence data.....42
3.3.2 Growth temperature.....42
3.3.3 Statistical analyses.....44
3.4 Results.....46
3.4.1 Nucleotide composition and sequence length of 16S rRNA in prokaryotes.....46
3.4.2 Genus level comparisons.....47
3.4.3 Phylogenetic-based comparison.....50
3.4.4 Nucleotide composition and length of vertebrates 18S rRNA.....54
3.5 Discussion.....56

Chapter 4 Mutational bias affects protein evolution in flowering plants.....61
4.1 Abstract.....61
4.2 Introduction.....62
4.3 Materials and Methods.....62
4.3.1 Sources of sequence data.....63
4.3.2 Identification and comparison of homologous sequences.....63
4.3.3 Identifying amino acids for GC-rich and AT-rich codons.....64

4.4 Results.....64
4.4.1 Compositional distribution of rice and *Arabidopsis* homologous genes.64
4.4.2 Amino acid substitutions between rice and *Arabidopsis* homologs.....67
4.4.3 Possible sources of compositional bias in rice genes and their encoded proteins.....72
4.4.4 Mutational bias affects protein sequence similarity.....75
4.5 Discussion.....76

Chapter 5 Nucleotide content affects synonymous codon usage in rice genes.....83
5.1 Abstract.....83
5.2 Introduction.....84
5.3 Materials and Methods.....85
5.3.1 Coding sequence data.....85
5.3.2 Identification of homologous sequences.....86
5.3.3 Statistical analyses.....86
5.3.3.1 G+C content and GC disparity.....86
5.3.3.2 Codon usage indices.....87
5.3.3.3 Correspondence analysis.....90
5.4 Results.....91
5.4.1 G+C content distribution and G+C disparity.....91
5.4.2 Relative synonymous codon usage.....92
5.4.3 Codon usage entropy.....96
5.4.4 Effective number of codons.....97
5.4.5 Correspondence analysis.....100
5.4.6 Codon adaptation index.....105
5.5 Discussion.....107

Chapter 6 Conclusions.....111
6.1 Thermal adaptation of rRNA genes and the genomes.....111
6.2 Effects of nucleotide bias on codon usage and protein evolution.....113

6.3 Future directions.....115
6.3.1 GC content, isochores and protein evolution in vertebrates.....115
6.3.2 Nucleotide bias, thermal adaptation and other forms of selection.....116
6.3.3 Other perspectives.....118

References.....121

List of Tables

TABLE 2.1 GC content and optimal growth temperature (T_{opt}, °C) of completely sequenced genomes used in this study.....25

TABLE 2.2 Average nucleotide composition (mean% ± standard error) of whole genomes and 16S rRNA genes of mesophiles and thermophiles. A) Nucleotide composition of entire genome, for the 31 genomes listed in Table 1. B) Nucleotide composition of 16S rRNA genes.....26

TABLE 2.3 Average nucleotide composition (mean% ± standard error) of structural components of 16S rRNA genes of 44 mesophiles and thermophiles.....27

TABLE 2.4 Correlation and regression analysis of nucleotide composition of 16S rRNA and optimal growth temperature.....31

TABLE 2.5 The relationship between the overall G+C content of the genome and the G+C content of A) 16S rRNA unpaired regions and B) the ribosomal protein gene coding sequences.....33

TABLE 3.1 Average GC content and sequence length of 16S rRNA stems and loops for mesophilic (< 40°C), moderately thermophilic (40-75 °C) and hyperthermophilic (≥ 75°C) bacteria (A) and archaea (B).....46

TABLE 3.2 Average 16S rRNA GC content (%) and length (bases) of mesophilic species and thermophilic species in the same genus.....48

TABLE 3.3 Average GC content (%) and cumulative length (bases) of stems and loops of 16S rRNA of mesophilic species and thermophilic species in the same genus.....49

TABLE 3.4 Correlation coefficients of GCrRNA and rRNA length with optimal temperature for original data (20 species), contrasts based on the translation tree (18 species) and contrasts based on the transcription tree (20 species).....50

TABLE 4.1 Average nucleotide contents of homologous genes in rice and *Arabidopsis* (expressed as percentages of G+C).....66

TABLE 4.2 Exon-intron structure of rice genes and their *Arabidopsis* homologs.....80

TABLE 5.1 A and B) Cumulative RSCU for 14005 rice and 25625 Arabidopsis genes. C1, C2, C3) RSCU for the three GC sets of rice genes.....94

TABLE 5.2 Gene number and average GC content (%) of high, intermediate and low GC

sets of rice genes and *Arabidopsis* homologs.....95

TABLE 5.3 Average and range of codon usage entropy (in bits) of high, intermediate and low GC sets of rice genes and *Arabidopsis* homologs.....97

List of Figures

FIGURE 1.1 GC content of ribosomal protein L3 gene. The species are separated as plants, invertebrates, vertebrates and fungi.....4

FIGURE 1.2 A phylogeny tree and the difference between directional and non-directional comparisons.....16

FIGURE 1.3 Two traits appear to be correlated in cross-species scatter plot (left figure). However, if the phylogeny of the species shows they form two clades, with the round species being in one clade and the triangle species in another clade (right figure) then there is no correlation of the traits within either clade.....17

FIGURE 2.1 G+C contents of 16S rRNA paired regions (stems) and unpaired regions (single strand regions) plotted against optimal growth temperature.....29

FIGURE 2.2 Individual nucleotide composition of unpaired regions of 16S rRNA genes plotted against optimal growth temperature.....30

FIGURE 2.3 G+C content of 16S rRNA paired and unpaired regions plotted against the average G+C content of the whole genome.....32

FIGURE 2.4 The relationship between the average genomic G+C content and the G+C content of (i) ribosomal protein genes and (ii) 16S rRNA unpaired regions.....33

FIGURE 2.5 A G+C plot of *M. jannaschii* genome.....36

FIGURE 3.1 Correlation of GC content of 16S rRNA stem regions and optimal growth temperature in 1573 bacterial species.....40

FIGURE 3.2 Distribution of optimal growth temperature of *Bacillus sp*.....43

FIGURE 3.3 Distribution of optimal growth temperature of 1673 prokaryotic species...44

FIGURE 3.4 Archaeal phylogenetic trees based on transcription proteins (a) and translation proteins (b).....45

FIGURE 3.5 Regression of GCrRNA contrasts on temperature contrasts based on the transcription tree of 20 archaeal species.....51

FIGURE 3.6 Regression of rRNA length contrasts on temperature contrasts based on the transcription tree of 20 archaeal species.....52

FIGURE 3.7 Regression of rRNA length contrasts on temperature contrasts based on the transcription tree, when the two *Methanosarcina* associated outlier points in

Figure 3.6 were removed.....52
FIGURE 3.8 Regression of GC rRNA contrasts on temperature contrasts based on the translation tree.....53

FIGURE 3.9 Regression of rRNA length contrasts on temperature contrasts based on the translation tree.....53

FIGURE 3.10 Average 18S rRNA G+C content and standard deviations (error bars) of 84 vertebrate in five groups.....54

FIGURE 3.11 Average 18S rRNA length and standard deviation (error bars) of 84 vertebrates in five groups.....55

FIGURE 3.12 Average nucleotide composition and standard deviations (error bars) of 18S rRNA stems of 84 vertebrate in five groups.....55

FIGURE 3.13 Average nucleotide composition and standard deviations (error bars) of 18S rRNA loops of 84 vertebrate in five groups.....56

FIGURE 3.14 The GC content of 16s rRNA stems increases rapidly with optimal growth temperature in 100 archaeal species, while average stem GC of 38 warm-blooded vertebrates have little increase over average stem GC of 46 cold-blooded vertebrates. The GC content of archaeal loops shows little increase with the temperature while the average loop GC of warm-blooded vertebrates is higher than that of cold-blooded vertebrates.....57

FIGURE 4.1 According to GC composition at first two codon positions, the universal codon table is partitioned into AU (AT)-rich codons, GC-rich codons and the other unbiased codons. The corresponding amino acids for the codons are shown in the four rectangle boxes, represented as single letters 64

FIGURE 4.2 Distribution of GC contents among rice and *Arabidopsis* genes. (A) 7886 rice genes and 25625 Arabidopsis genes. (B) 4447 homologous gene pairs of rice and *Arabidopsis* homologs.....65

FIGURE 4.3 Amino acid content of High GC rice and homologous *Arabidopsis* protein sequences. (A) The content of G,A,R,P and F,Y,M,I,N,K amino acids for High GC rice genes and their *Arabidopsis* homologs. (B) Proportions of individual amino acids at variant sites plotted for the High GC rice genes and their homologs from *Arabidopsis*.....68

FIGURE 4.4 Average amino acid contents of Low GC rice and *Arabidopsis* homologous protein sequences at variant sites.....69

FIGURE 4.5 Amino acid exchange matrix for High GC rice gene encoded proteins and their *Arabidopsis* homologs.....70

FIGURE 4.6 Amino acid exchange matrix for Low GC rice gene encoded proteins and their *Arabidopsis* homologs.....72

FIGURE 4.7 GC content plots along the coding sequence for rice high GC, low GC genes (4447 genes in total) and their *Arabidopsis* homologs (4447 genes in total).....73

FIGURE 4.8 The degree of mutational bias correlates with position within the gene.....73

FIGURE 4.9 The relationship between coding sequence length and GC content.....74

FIGURE 4.10 The correlation between sequence divergence and the GC content of rice homologs.....76

FIGURE 4.11 dN and dS in rice High GC genes and Low GC genes compared to *Arabidopsis* homologs.....78

FIGURE 4.12 Histogram of dN/dS distribution in the rice low GC and high GC gene sets compared to *Arabidopsis* homologs.....78

FIGURE 4.13 Average evolutionary distance of Low GC and High GC rice genes and *Arabidopsis* homologs, calculated using Tamura & Nei (1993) method and Tamura & Kumar (2002) method.....79

FIGURE 5.1 The G+C content of 14005 rice genes shows a two modal distribution and was separated into low, intermediate and high GC classes.....91

FIGURE 5.2 The GC content disparity plot for rice genes. A) GC3 > GC1 > GC2 in high GC genes. B) also the pattern GC3 > GC1 > GC2 is seen in intermediate GC genes. C) GC1 > GC3 > GC2 in low GC genes.....93

FIGURE 5.3 GC content of high, intermediate and low GC sets of rice genes plotted against *Arabidopsis* homologs.....96

FIGURE 5.4 Nc plot for 14005 rice genes with ribosomal protein genes highlighted.....98

FIGURE 5.5 Nc plot for 25625 *Arabidopsis* genes.....99

FIGURE 5.6 Nc plot for 7160 rice genes homologous to *Arabidopsis* genes.....99

FIGURE 5.7 Nc plot for 7160 *Arabidopsis* genes homologous to rice genes.....100

FIGURE 5.8 Correspondence analysis of RSCU of 14005 rice genes. A) the distribution of the genes. B) the distribution of the codons.....101

FIGURE 5.9 Correlation between GC1, GC2 and GC3 with gene locations on the

primary axis of the correspondence analysis in Figure 5.8A.....102

FIGURE 5.10 Correspondence analysis of relative synonymous codon usage of 7160 rice genes and 7160 *Arabidopsis* homologs. A) gene distribution. B) codon distribution.....103

FIGURE 5.11 Correlation of G+C content of 7160 rice genes and 7160 *Arabidopsis* homologs with their positions on the primary axis of the correspondence analysis 104

FIGURE 5.12 Correlation of GC1 and GC2 of 14320 rice and *Arabidopsis* homologous genes with their positions on the primary axis of the correspondence analysis 104

FIGURE 5.13 GC3 of 14005 rice genes including ribosomal protein genes are linearly correlated with CAI.....106

FIGURE 5.14 GC1 and GC2 of 14005 rice genes show weak positive correlation with CAI.....106

FIGURE 5.15 Correlation of GC3 and CAI (at iteration 7) in 6627 Arabidopsis chromosome 1 genes.....107

FIGURE 5.16 Correlation of frequency of GC1 and GC2 in coding regions and GC frequency in introns, respectively, with GC3s for 9620 rice genes having an intron length greater than 500 bases.....109

FIGURE 6.1 Molecular response to thermal adaptation.....113

Chapter 1

General introduction

The genome may be defined as all of the genetic material (mainly the DNA) contained in the cell. DNA consists of four types of nucleotide bases: adenine (A), cytosine (C), guanine (G) and thymine (T). Quantitative analyses of the composition of these four bases (the molar amount of the A, C, G and T) were pioneered by Erwin Chargaff in the late nineteen forties, before the double helix structure of DNA was proposed. Chargaff (1950) noted the following DNA compositional regularities:

(1) **A** + **G** = **T** + **C**, *i.e.*, the amount of the purines equals that of the pyrimidines;

(2) **A** = **T**;

(3) **G** = **C**; and as a logical consequence of these three equations:

(4) **A** + **C** = **G** + **T**, *i.e.*, the sum of the 6-amino compounds equals that of the 6-oxo derivatives.

These equations are consistent with the base pairing rules of DNA double helix structure. In fact, they actually led to the double helix proposal of Watson and Crick (1953). Chargaff further found that the base parity rule of **A** = **T** and **G** = **C** also applies approximately to single strands of DNA, which indicates that DNA single strand can form local helical structure (Chargaff, 1979; Forsdyke & Mortimer, 2000). Another important finding made by Chargaff (1950) was that the composition of the four nucleotides varies from species to species, but not from one tissue to another within a given species. This demonstrated the nucleotide composition, usually measured as molar fraction of guanine plus cytosine (G+C or GC content) in a genome, is a characteristic of a species.

1.1 Variations (bias) in GC content

As more and more GC contents of different genomes were determined, it became clear that nucleotide composition between genomes (species) can be very different (Lee et al., 1956; Belozersky & Spirin, 1958; Lobry & Sueoka, 2002). Some organisms have genomes that are disproportionately rich in G and C, whereas others have genomes rich

in A and T. Bacterial species have the widest range of GC content, from 25% in *Mycoplasma capricolum* to over 75% in the gram-positive actinobacterium *Micrococcus luteus* (Li, 1997, p.401; p.404). Protist and algal genomes also have large variations in GC content. The genome of *Plasmodium falciparum*, for example, is extremely GC poor (18%), while GC content of *Giardia lamblia* is 46%. The range of average genomic GC contents diminishes within both plants and animals, and vertebrate GC content ranges from 37% in the tuna *Thunnus thunnus* to 50% in the lamprey *Lampetra fluviatilis* (Mooers & Holmes, 2000). Overall, vertebrate genomes tend to be GC poor. Our human genome, for example, has an average GC content of 41% (The Genome Sequencing Consortium, 2001).

While intergenomic variation of GC content is large in bacteria, intragenomic heterogeneity is generally small (Rolfe & Meselson, 1959; Sueoka et al., 1959; Muto & Osawa, 1987), although there are some exceptions (Kerr et al., 1997; Sueoka, 1999). The GC content of some vertebrate genomes, however, is characterized by a relatively large intragenomic variability in base composition (Sueoka, 1962; Bernardi et al. 1985). Neglecting satellite DNAs, the genomes of warm-blooded vertebrates cover a very broad compositional spectrum (30-60% GC) (Alvarez-Valin et al., 2002). Variability is usually seen among different regions of the chromosomes and they are discontinuous. It is likely that human and other warm-blooded vertebrate genomes are mosaics of fairly large regions of similar base composition called "isochores" which are generally more than 200 kilobase pairs in length (Bernardi et al. 1985). But even the GC content within an isochore is not compositionally homogeneous, as the whole human genome revealed (The Genome Sequencing Consortium, 2001). In contrast, the genomes of cold-blooded vertebrates are characterized by a much lower GC heterogeneity and lack GC-rich isochores.

1.2 Effects of GC variations

It has been found that the GC content of structural components of many genomes is correlated with GC content of the entire genome, particularly in bacteria. This was first suggested by Sueoka (1961a), who pointed out that GC content of individual strains of *Tetrahymena* tends to be uniform throughout the genome. When the GC content of a

particular strain changes, all the molecules undergo increases or decreases of GC pairs in similar amounts. This observation has been abundantly confirmed for a variety of species (Muto & Osawa, 1987). The bacterial genome is roughly composed of protein coding genes (70-80%), spacers including various signals (20-30%) and structural RNA genes (<1%). All components are positively correlated with genomic GC content, although the correlation coefficients are different. Spacers and protein genes have the strongest linear correlation whereas transfer RNA and ribosomal RNA genes have weaker correlation. Within protein genes, the GC content of all three codon positions are correlated with genomic GC content, with GC content at third position being the strongest, followed by that at first position and second position being the weakest (Bernardi & Bernardi, 1986; Muto & Osawa, 1987; Wilquet & Van de Casteele, 1999). Therefore it is not unexpected that GC content will directly influence the codon usage and amino acid usage in these organisms.

Variation in genomic nucleotide composition is also accompanied by subtle but significant shifts in the amino acid composition of proteins. This was originally suggested in Sueoka (1961b), even before the genetic code was deciphered. He reported that the amino acid composition of total bacterial proteins in GC rich organisms is different from that in GC poor organisms. According to correlation between GC content and amino acid composition, he classified amino acids into three groups: alanine, arginine, glycine and proline are positively correlated with the GC content; isoleucine, lysine, aspartic acid plus asparagines, glutamic acid plus glutamine, tyrosine and phenylalanine are negatively correlated; histidine, valine, leucine, threonine, serine and possibly methionine are extremely uniform with no detectable evidence of correlation (Sueoka, 1961b). More recently, various studies have compared homologous protein sequences of GC-rich and GC-poor organisms, showing a positive correlation between genomic GC content and amino acid usage: GC rich organisms have a higher composition of amino acids encoded by GC-rich codons and vice versa (Collins & Jukes, 1993; Foster, Jermiin & Hickey, 1997; Lobry, 1997; Nakachi et al., 1997; Gu, Hewett-Emmett & Li, 1998; Wilquet & Van de Casteele, 1999; Singer & Hickey, 2000). Some exceptions to the rule include elongation factor (EF)-2, EF-1alpha, the largest subunit of RNA polymerase III, glyceraldehydes 3-phosphate dehydrogenase, and mitochondrially encoded proteins.

Nucleotide composition also has phylogenetic implications. On one hand DNA base composition is a reflection of phylogenetic relationship (Sueoka, 1961a). Closely related species have similar GC content. For example Figure 1.1 shows GC content of ribosomal protein L3 gene from 18 eukaryotic species. The two monocot plants (rice and

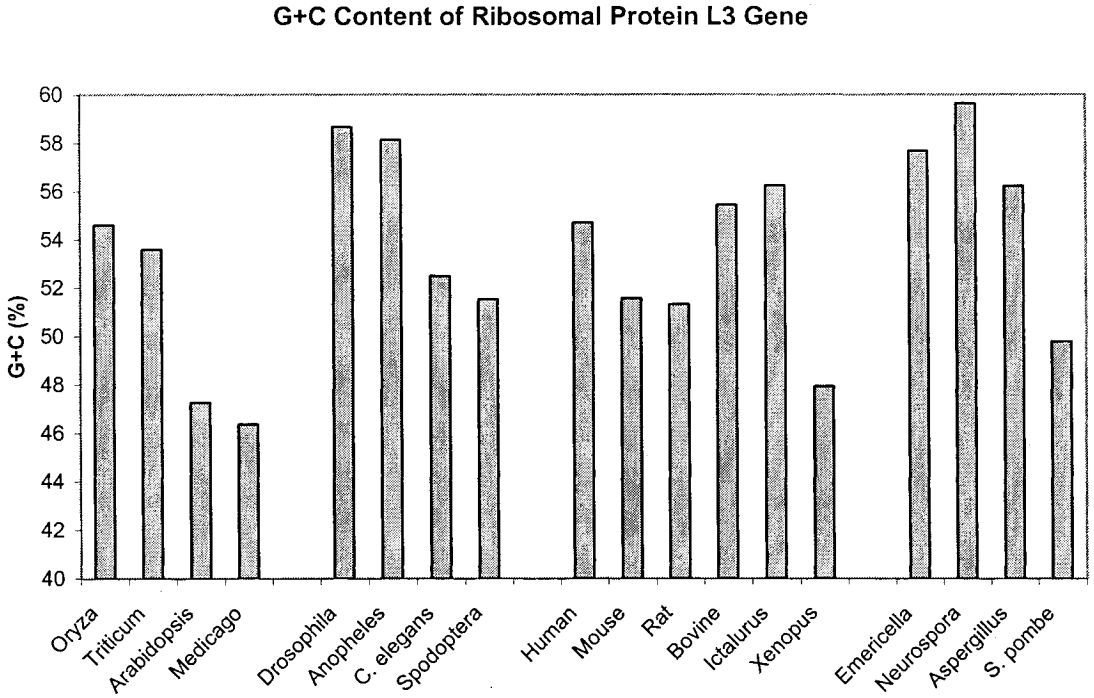


Fig. 1.1 GC content of ribosomal protein L3 gene. The species are separated as plants, invertebrates, vertebrates and fungi.

Triticum) have similar GC contents and likewise the two dicot plants (*Arabidopsis* and *Medicago*). In the past, when DNA sequences were scarce, GC content has been used for inferring phylogenetic trees of bacteria (e.g., Muto & Osawa, 1987). On the other hand nucleotide bias, including GC content bias and codon bias, can affect phylogenetic reconstruction based on DNA sequences (Foster, 1997). For example, Hasegawa and Hashimoto (1993) found phylogenetic analyses based on ribosomal RNA gene sequences could be unreliable due to extreme GC content bias in the rRNA genes of some taxa. This prompted people to favor the use of protein sequences for inferring phylogenetic trees, as nucleotide bias in protein-coding genes are mainly reflected in the third codon positions, which usually do not affect encoded amino acid. However, previous work in our lab has

shown that biased GC content can lead to biased amino acid composition, which can result in erroneous phylogenetic reconstruction based on protein sequences (Foster, 1997; Foster & Hickey, 1999).

1.3 Biased DNA mutation and natural selection shape GC content

For the four DNA nucleotides there are 12 possible substitutions, that is A ↔ C, A ↔ G, A ↔ T, C ↔ G, C ↔ T and G ↔ T (the ↔ sign means the substitutions can proceed in two directions, for instance A ↔ C means either A → C or C → A). If all substitutions occur randomly, i.e., there is no bias in the direction of change so that the rate of the 12 substitutions are all equal, as assumed in Jukes and Cantor's (1969) one-parameter model, then a random sequence will have equal amount of A, C, G and T at equilibrium, which leads to a GC content of 50%. Kimura's (1980) two-parameter model in which the rate of transitions (purine changing to purine and pyrimidine changing to pyrimidine) is different from the rate of transversions (purine changing to pyrimidine, or *vice versa*), also predicts the equilibrium value for the frequency of each nucleotide to be ¼ and thus the GC content will also be 50%. However, neither of the two assumptions is realistic because, as we have seen in Section 1.1, in many cases the GC content of a genome and its individual genes deviate significantly from 50%.

For instance, among 13 mammalian pseudogene sequences, the rate of each pair of mutual mutations is different (Li & Graur, 1991, p.90). For example the rate of A → G mutation is 9.4% while that of G → A mutation is 20.7%; the rate of A→ C mutation is 5% and that of C→ A is 6.5%; the rate of A→ T mutation is 4.7% and that of T→ A mutation is 4.4%. Rates of transitions are higher than rates of transversions. As we can see here A→ G mutation is almost the summation of A→ C and A→ T mutations, and G→ A mutation is much higher than the summation of C→ A and T→ A. Overall, the data indicate that there are more C and G to A and T mutations (55.4%) than A and T to C and G mutations (25.9%); consequently, mammalian pseudogenes (and other non-selected sequences) tend to be AT-rich.

1.3.1 Selectionist interpretations of GC content change

Two main hypotheses have been proposed to explain the variation in GC contents: natural selection and neutral evolution. In the selectionist view, GC content is a form of adaptation to environmental conditions and a high GC content has often been suggested to be advantageous, because the GC pair has three hydrogen bonds and the AT pair has only two, so a high-GC DNA will be more stable than a high AT DNA (Wada & Suyama, 1986). One scenario for natural selection is the ultraviolet sensitivity of unicellular organisms. Singer and Ames (1970) found in general the higher GC organisms were those deemed to have a high UV-exposure, and the lower GC organisms were the others. The high UV-exposure species were largely those which reproduced and/or dwelled aerially, at or very near the soil surface or within the surface layers of water bodies. Organisms with intermediate UV exposure often showed accordingly intermediate GC content. For example, *E. coli* has a life cycle that split into enteric (gut-dwelling) and external periods and its GC content is 50%. These findings let Singer and Ames (1970) propose that UV-exposure of a species was positively correlated with its GC content.

Another selectionist scenario invokes thermophily as a selection force. Thermophilic bacteria show strong preferential usage of thermally stable amino acids encoded by GC-rich codons (such as alanine and arginine) and strong avoidance of thermally unstable amino acids encoded by GC-poor codons (such as serine and lysine) (Li, 1997, p.401). Haney et al. (1999) compared mesophilic and thermophilic *Methanococcus* species and found indeed three (R, P and A) of the four GC-rich codon encoded amino acids (G, A, R and P) have higher frequency in *M. jannaschii*. Serine is very low in *M. jannaschii*. Two amino acids (N and M) of the six AT-rich codon encoded amino acids (F, Y, M, I, N, K) have lower frequency in *M. jannaschii*. Phenylalanine (F) has almost same frequency in *M. jannaschii* and other mesophilic *Methanococcus* species. But these amino acid composition differences in thermophiles and mesophiles were not demonstrated in the study of Singer and Hickey (2003). For instance, they found alanine is significantly higher in the mesophiles, contrary to the above results.

In vertebrates, Bernardi (1985) found that warm-blooded birds and mammals have a higher GC content than cold-blooded reptiles and amphibians. However, no correlations between GC content of the genome or at the third codon position of the

coding sequence and growth temperature has been found (Dalgaard JZ & Garrett RA. 1993; Galtier & Lobry, 1997; Hurst & Merchant, 2001; Wang & Hickey, 2002; Belle, Smith & Eyre-Walker, 2002), which suggests that the GC content is not controlled by the temperature. Therefore, there are ongoing debates on the relationship of GC content and thermal adaptation, which is one focus of this thesis. This will be elaborated in Chapters 2 and 3.

1.3.2 Neutralist interpretations of genomic differences in GC content

The neutralist view invokes biases in DNA mutation patterns to explain the variation in GC content. When an organism has a tendency to change GC base pairs into ATs or *vice versa*, mutations are said to be biased. The ratio of the two tendencies is called mutation pressure; since the two tendencies are biased it is called biased mutation pressure. For intragenomic variation in GC content, a neutral interpretation is that biased mutation pressure has been exerted uniformly or near-uniformly on all parts of the genome, but is counterbalanced by region-specific selective constraints.

A bacterial genome consists primarily of protein-coding regions, non-coding spacer regions and structural RNA genes such as tRNA and rRNA genes. Each of these parts tends to show a similarly-biased GC content; that is if a genome is GC-rich, all parts of the genome will be GC-rich. However, the extent of the GC-richness may be different. For instance, spacer DNAs show the greatest GC bias. Protein coding regions have slightly reduced GC imbalance compared to the spacers. Structural RNA genes will also show some nucleotide imbalance but less than the overall average genomic nucleotide bias. Within protein coding genes, all three codon positions will show similar GC imbalance, but GC3 (GC content at 3rd codon position) will show greatest variation, followed by GC1 and GC2 has the least GC imbalance. These results support and are explained by the neutral theory of molecular evolution: biased mutation pressure exerts uniformly on all parts of the genome and the variation in GC content is caused by differences in functional constraint between the different regions. Spacer regions and the GC3 of protein coding genes have little functional constraint therefore mutations are mostly neutral in effect and have the greatest chance of surviving, thus causing a big GC imbalance. Structural RNA genes and GC2 have the greatest functional constraints so

most mutations will be deleterious and selected against from the outset, therefore only a small portion of mutational bias will be shown in them. Consequently, GC variation will be smaller in these regions. Indeed, the overall genomic GC content of 529 prokaryotic species varies between 23% and 77% (data based on Galtier & Lobry, 1997), while the range of the corresponding 16S rRNA GC content is only 45 to 69%.

The neutral interpretation presented above addresses intragenomic variation in GC content, which is generally small. It does not, however, explain the intergenomic variation, which is large in bacteria. The large GC variation between different bacterial species was addressed by Noboru Sueoka (Sueoka, 1962; 1988). Mutational bias is essentially the conversion of GC pairs to AT pairs or *vice versa*. Sueoka (1962) made a major assumption that the rate of conversion was more or less equal between different base pairs. Let u be the rate of substitution from G/C to A/T and v be the rate of substitution from A/T to G/C. The u and v values are the probabilities per base pair per generation of an *effective base conversion* taking place, which means these mutations survive into the next generation. Suppose the GC content of a genome at a generation $\boldsymbol{n}$ is π_n, and the AT content is then $(1-\pi_n)$, then at generation $(n+1)$ the GC content will be

$$\pi_{n+1} = \pi_n + v(1-\pi_n) - u\pi_n \qquad \textbf{(2.1)}$$

Or the change in π_n, $\Delta\pi$ is:

$$\Delta\pi = v(1-\pi_n) - u\pi_n = v - (u+v)\,\pi_n \qquad \textbf{(2.2)}$$

At equilibrium $\Delta\pi = 0$, so that the genomic GC content is

$$\boldsymbol{\pi_{GC} = \pi_n \text{ (equilibrium)} = v/(v+u)} \qquad \textbf{(2.3)}$$

Here we can see the GC content is only determined by the ratio v/u. When it is 1.0 then at equilibrium GC content is 50%, as in Escherichia coli. When the ratio is 3.0 GC will be 25%, as in *Mycoplasma capricolum*. When it is 0.33, GC will be 75%, as in Micrococcus luteus (Sueoka, 1962; Li, 1997, p. 401).

The ratio $\boldsymbol{v/(v+u)}$ is also called biased mutation pressure, μ_D (Sueoka, 1988). A $\mu_D > 0.5$ indicates a bias towards GC pairs and a $\mu_D < 0.5$ indicates a bias towards AT pairs. Therefore, mutation biases have the ability to cause a directional pattern of evolution of GC content. DNA mutation and mutation repair are two main mechanisms that cause mutation bias. Any change in mutation and repair processes, together with the biases in

the fixation of mutated bases in a population, will bring about GC content change. In the following section we will discuss sources and molecular mechanisms of biased mutation and biased repair.

1.3.3 DNA mutation

DNA mutations can be induced by exposure to exogenous mutagenic factors, or they can occur spontaneously in the absence of such exposure. As DNA replicates, the incorrect base is sometimes inserted into the growing DNA chain. For instance, in *Escherichia coli* replication, for every 10^{10} nucleotides incorporated there are about 5.4 mistakes. Yeast, mouse and human have a smaller mutation rates ($2.2*10^{-10}$, $1.8*10^{-10}$ and $5*10^{-11}$, respectively) (Drake et al., 1998). A common type of mismatch is the incorrect incorporation of a T opposite a G. After the first replication, the mispaired T will pair with an A, causing a GC-to-AT change in the sequence of one of the two progeny DNAs and thus changing the base pair at that position on all subsequent copies of the mutated DNA molecule.

Other mutations during DNA replication include dislocation (or misalignment) of the primer-template and replication slippage or slipped-strand mispairing, which occurs when replicate DNA regions contain contiguous short repeats, and can result in either deletion or duplication of a DNA segment (Li, 1997, pp.26-29).

In addition to mutations in DNA replication processes, errors can occur in DNA recombination and gene conversion. Insertion, deletion and inversions of DNA fragments during recombination events can cause chromosomal structure change and thus produce large GC content change of the local chromosomal region. Gene conversion is said to be biased if the probability of conversion is not symmetric (e.g., if GC alleles convert AT alleles more frequently than the reverse). It has been suggested that the direction of the conversion might be biased toward G and C in mammals. Biased gene conversion has recently been proposed as a new mechanism to explain the variation in GC content along mammalian chromosomes (i.e., isochores) (Galtier et al., 2001; Eyre-Walker & Hurst, 2001; Duret et al., 2002).

DNA strands are often methylated. This is a mechanism that bacteria use to protect their chromosomal DNA from endonucleases which cleave foreign DNAs such as

bacteriophage. In eukaryotes DNA methylation is involved in gene regulation. Cytosines of higher organism genomes are often methylated to be 5-methylcytosine (m^5C). About 80% of CpG occur in the methylated form (on both strands). Strikingly, the total occurrence of CpG (methylated or not) is only 20% of that expected from simple binomial statistics based on the frequency of occurrence of the four bases in DNA, *i.e.*, $P_{CpG}/P_cP_G = 0.2$. The remainder of the expected CpG has apparently been lost through mutation. Indeed, CpG is a mutation hotspot in the human genome. The occurrence of the product of the mutation, TpG, from deamination of m^5C in the CpG is elevated, as expected. This is a remarkable example of how a small bias, over the evolutionary time scale, can lead to dramatic alteration in base composition.

In addition to the methylation of cytosines, several DNA bases can be altered by deamination. For example, when A is deaminated it becomes hypoxanthine, which pairs with C during replication, incorporating C instead of T at that position. In a subsequent replication, the C will pair with G, causing an AT-to-GC transition in the DNA. When C is deaminated, it becomes uracil (U). U will pair with A during replication, causing a GC-to-AT transition. When G is deaminated, it becomes xanthine. Since xanthine also pairs with C, this will restore to correct GC pair in subsequent replication.

Induced DNA mutations bring about DNA lesions or increase the rate of spontaneous mutations. Heat can speed up spontaneous chemical reaction, leading, for example, to the deamination of bases. Chemicals such as deaminating agents like hydroxylamine, bisulfite and nitrous acid can greatly increase the rate of deamination. Deamination is mutagenic because it results in base mispairing. Alkylating agents such as ethyl methanesulfonate and methyl methanesulfonate add alkyl groups (CH_3, CH_3CH_2, etc.) to the bases or phosphates. For example, alkylation of guanine produces O^6-methylguanine. The altered base sometimes pairs with thymine, causing mutations. Reactive oxygen can oxidize guanine to 7,8-dihydro-8-oxoguanine (8-oxoG), which mispairs with adenine in DNA synthesis.

Ultraviolet irradiation due to sun exposure can cause thymine dimers (TT). We already see that UV sensitivity is an argument for the selectionist view of GC content change. Singer and Ames (1970) suggested that a DNA with 75% GC content has 60% fewer photodimerization targets (i.e., TT dimers) than DNA with 25% GC. A simple

calculation based on base frequencies shows an even greater difference in target frequency: at 75% GC, the probability of any given base being a thymine is 0.125, so the chance of finding two adjacent thymines is 0.125 * 0.125 = 0.015625. At 25% GC the probability of any given base being a thymine is 0.375, so the chance of finding two adjacent thymines is 0.375 * 0.375 = 0.140625. This is an almost ten-fold difference from the value at 25% GC (http://www.btinternet.com/~neil.dec/fot-mine/fot-mine-diss/fot-diss-evo.htm).

1.3.4 Bias in mutation

As noted above, there are more GC to AT mutations than AT to GC mutations in mammalian pseudogenes (Li & Graur, 1991, p.90). Duret *et al.* (2002) found an excess of GC → AT over AT → GC changes in synonymous sites of mammalian genes, especially in GC-rich genes. Both pseudogene and synonymous sites of protein coding genes are subject to little functional constraint, therefore the bias reflects AT mutational pressure in mammalian genomes. Further studies found in addition to human and mouse, many other organisms including *Drosophila melanogaster*, yeast, *zea maize* and *E. coli* are subject to this AT-biased mutational pressure (Birdsell, 2002; Alvarez-Valin, Lamolle & Bernardi, 2002).

Such mutational pressures can result from a variety of chemical processes mentioned above, including cytosine and 5-methycytosine deamination, oxidative damage to cytosine or guanine, or UV irradiation, all of which can result in GC to AT or TA mutations. Furthermore, different DNA polymerases have different mispairing frequencies for different bases (Kunkel & Alexander, 1986). For example, for base A, the number of mis-substitutions by the other bases in replication is in the order T>G>>C; for C, mis-substitutions are in the order T>>A>G. Both cause AT-biased mutations.

Another source of mutation bias is the bias in the free nucleotide pool of a cell. If a mis-incorporation is going to occur, then if there are more, say, adenines and thymine available then cytosine and guanine, then it is more likely an A or a T will be mis-incorporated. Free nucleotide concentrations vary during cell cycle, so regions of the genome that replicates at different times should have different mutation pattern (Eyre-Walker & Hurst, 2001).

Given constant mutations in DNA and with a ubiquitous AT-biased mutation pattern, the cell must have mutation repair mechanisms to repair DNA mismatches and it is reasonable to believe repair bias, if any, will bias toward GC over AT so that the genome will not become too AT rich.

1.3.5 DNA mutation repair and repair bias

The cell has many types of repair systems to restore mutated bases or repair damaged DNA, depending on the types and degree of the damage. For instance, methyl-directed mismatch repair system corrects mismatch in DNA replication. In 5-methylcytosine induced mutation, 5-methylcytosine deaminates and becomes a thymine. The latter cannot be recognized as an unusual base and thus cannot be removed by uracil-N-glycosylase. This will cause a G/T mismatch in the DNA. The cell has a mechanism called very short patch (VSP) repair that recognizes and cuts the mismatched T and inserts the correct base C in the DNA strand. However, if T is really a correct base, this will cause an AT-to-GC mutation. Hence the VSP system is GC-biased repair.

In a study of repair of the 12 single-base mismatches in recombination intermediates in Chinese hamster ovary cells, Bill et al. (1998) found there was significant repair of G-T $\rightarrow$ G-C, with a slightly greater bias in a CpG context. Repair of C-A was also biased (toward C-G). Among the heteromismatches, the trend was toward retention of C or G vs. A or T. This study suggested the mismatch repair systems tend to increase the GC content. It was confirmed by a large scale analysis of base mismatch repairs (Birdsell, 2002). Of 72971 repaired heteromismatches collected in 50 experiments in *S. cerevisiae*, 42242 (57.9%) were repaired to G/C or C/G, whereas only 30729 (42.1%) were repaired to A/T or T/A. The mean ratio of repair to GC versus AT for the 50 experiments is 1.48 to 1. The GC repair bias was most pronounced for G/T mismatches which exhibited a mean bias of 1.71 to 1, followed by C/T mismatches (1.5 to 1), A/G mismatches (1.38 to 1) and A/C mismatches (1.31 to 1).

1.3.6 Interaction of neutral evolution and selection

We have seen that there is a widespread AT-biased mutation pressure and a GC-biased repair system to counter the effect of the mutation pressure. Depending on the balance of

the two forces the organism can be GC-rich or AT-rich. The mutational pressure may be exerted uniformly or near-uniformly on the whole genome, as the neutralist view proposes, so the difference in GC content among different regions of the genome is controlled by the function of the regions, where in most cases negative, purifying selection forces play an essential role. Those regions having little functional constraint (*e.g.*, spacer DNA sequences, pseudogenes, and the third codon positions of protein coding genes) show large GC content variation, whereas regions having great functional constraint (*e.g.* rRNA and tRNA genes, the first and second codon positions of protein coding genes) have small GC content variation. In Chapters 2 and 3 we will further investigate the relationship among environmental selection (growth temperature), functional constraint (helical stem regions and single stranded loop regions), and GC change in rRNA genes in prokaryotes and some eukaryotes (vertebrates).

In summary, biased DNA mutation, or directional DNA evolution, affects GC content, which will in turn affect codon usage and amino acid usage (see section 1.2). We will investigate this effect of GC content change on codon usage and protein evolution by doing a comparative study of homologous protein sequences of rice and *Arabidopsis* (see Chapters 5 and 4). It has been shown that the degree to which each protein is affected, depended on the degree to which natural selection conserves its sequence: highly conserved proteins were affected to a lesser degree than loosely conserved proteins (Singer, 2002). This means that DNA mutation bias at the level of the gene must interact with natural selection acting on the protein. This has been tested on the evolution of mitochondrial proteins of helical transmembrane domains. The results show that rather than acting on the amino acid sequence level, natural selection is conserving the transmembrane domains at the biochemical level by keeping hydrophobic residues, leaving some flexibility in their amino acid compositions, where DNA mutation bias can exert its effect (Singer, 2002). This indicated an interaction between directional neutral evolution and negative selection shapes the amino acid composition of the helical transmembrane domains. Furthermore, an interaction between neutral evolution and positive selection is also possible, where mutational bias acts on those positively selected sites to fine tune codon usage and amino acid usage. Singer (2002) applied the adaptive

landscape model (Wright, 1932) to illustrate this but further studies needed to confirm this interaction.

1.4 Research proposal

As demonstrated above, genomic nucleotide composition plays a fundamental role in genome structure, codon usage and amino acid usage, and therefore affects both gene/protein structure and genome/protein evolution. The causes and consequences of changes in nucleotide composition are increasingly recognized as being of major importance in genome evolution. Not only do they provide a valuable window on fundamental evolutionary processes, but also they greatly affect our ability to accurately reconstruct phylogenetic history (Mooers & Holmes, 2000).

The recent developments in whole genome sequencing projects have provided complete sequence data for over one hundred prokaryotic and eukaryotic genomes. Comparative analyses of nucleotide bias at the whole genome level and levels of various genomic components are now available. In this research, we have used whole genome information to study the forces that cause genomic nucleotide bias and its consequences on genome evolution. The objectives of this research are the understanding of evolutionary forces that drive nucleotide bias and understanding how this bias has led to amino acid sequence changes and genome evolution. The underlying logic is that by genome-wide sequence comparison we would be able to find patterns of nucleotide and protein bias in various organisms, which helps understand the process of genome evolution. The biological hypotheses include

(1) Nucleotide bias is the cause, rather than the consequences of protein change;

(2) Both mutation and selection forces cause nucleotide bias.

Corresponding to these two points, we have studied genomic nucleotide bias in two ways, one is the set of forces that drive nucleotide bias in different organisms; and the other is the consequence of nucleotide bias on genome evolution.

For the first task, as shown in Chapter 1.2 two main evolutionary processes have been invoked to explain why patterns of nucleotide composition vary within and among species: biases in the process of mutation such that the rates of changes from G•C $\leftrightarrow$ A•T are not constant in time or space; and natural selection, either on overall GC content

or on specific patterns of codon usage (Mooers & Holmes, 2000). We have specifically focused on the effect of natural selection on GC content. Thermal adaptation (the GC content is a response to environmental temperature) is a primary example to study the effect of natural selection. We will compare GC content difference between mesophilic and thermophilic bacteria/archaea and between warm-blooded and cold-blooded vertebrates.

For the second task, the selectionist would predict GC bias will have no effect on protein evolution of the genome, whereas the neutralist would predict GC bias does cause protein sequence change. Previous studies in our lab have found that nucleotide bias affects amino acid content in proteins coded by animal mitochondria (Foster, 1997; Foster, Jermiin & Hickey, 1997) and amino acid content in proteins coded by whole bacterial and yeast nuclear genomes (Singer, 2002; Singer & Hickey, 2000). This effect is in a predictable direction, whereby genes of high GC content encode proteins consisting of high frequency of amino acids of GC-rich codons. It will be very interesting to see whether the same pattern applies to multicellular organisms. As more and more complete sequences of eukaryotic genomes are available we can extend these studies to multicellular eukaryotes. For this we have compared GC content, codon and amino acid pattern in two flowering plants, *Oryza sativa* (rice) and *Aribidopsis thaliana*.

1.5 Comparative methods

Felsenstein (2004) pointed out: "Comparative methods use the distribution of traits across species to make inference about the effect on their evolution of other traits or of environments." A number of comparative studies are involved in this thesis work. To study GC content bias and thermal adaptation we will compare traits (*i.e.*, GC content) against optimal growth temperature of prokaryotes and compare traits of vertebrates in different temperature groups. To study the effect of GC content change on protein evolution we will compare GC content, codon usage and amino acid usage of homologous genes and proteins between rice and *Aribidopsis*. There are two kinds of comparative methods, cross-species comparison (non-directional comparison) and phylogeny-based comparison (directional comparison) (Fig. 1.2; Harvey & Pagel, 1991).

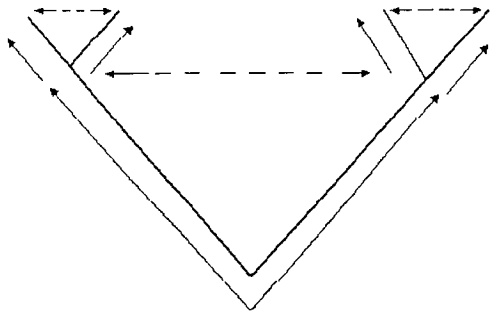


Fig.1.2 A phylogeny tree and the difference between directional and non-directional comparisons. The solid arrowed lines represent directional comparisons between ancestors and descendants, while the dotted arrowed lines represent non-directional comparisons between species or taxa at the same phylogenetic level. From Harvey & Pagel, 1991.

The non-directional cross-species comparison analyses evolutionary trends across either contemporary species, or across higher nodes which are usually at a similar taxonomic or phylogenetic level. It is a traditional way of comparative technique favored by Darwin and has been used by almost all evolutionary biologists ever since his time. Indeed studies using this approach have taught us most of what we know about adaptation. The phylogenetic-based directional comparison, developed in the last 25 years, makes use of phylogeny to infer the direction and rates of evolutionary change between ancestors and descendants. Now it is generally accepted that statistical analyses of interspecific data should be conducted in a phylogenetic context using phylogenetic comparative method (Martins & Housworth, 2002).

Why does phylogeny matter? The reason is that species are part of a hierarchically structured phylogeny, and thus cannot be regarded as independent samples from the same distribution. Similarity between species can vary according to their level of relatedness, rather than randomly, as would be expected if species traits were independently acquired (Felsenstein, 1985; Harvey & Pagel, 1991). Cross-species correlations of traits which do not account for this nonindependence are subject to inflated type I and type II errors and may be misleading (Martins & Garland, 1991; Grafen & Ridley, 1996; Harvey & Rambaut, 1998). For example Fig. 1.3 is a scatter plot of two traits for 60 species (represented as the points). It appears that the two traits are correlated – a large value of trait 2 goes with a large value of trait 1 (Fig.1.3 left). However, if the phylogeny of the 60 species is such that all round species are in one clade

and all triangle species are in another clade then there is no correlation within the two clades (Fig.1.3 right). The correlation between the two traits among the 60 species simply arises from the difference between the clades (Felsenstein, 2004). This example demonstrates a correlation between two traits is suggested by a cross-species comparison while it is not verified by the comparison in a phylogenetic context.

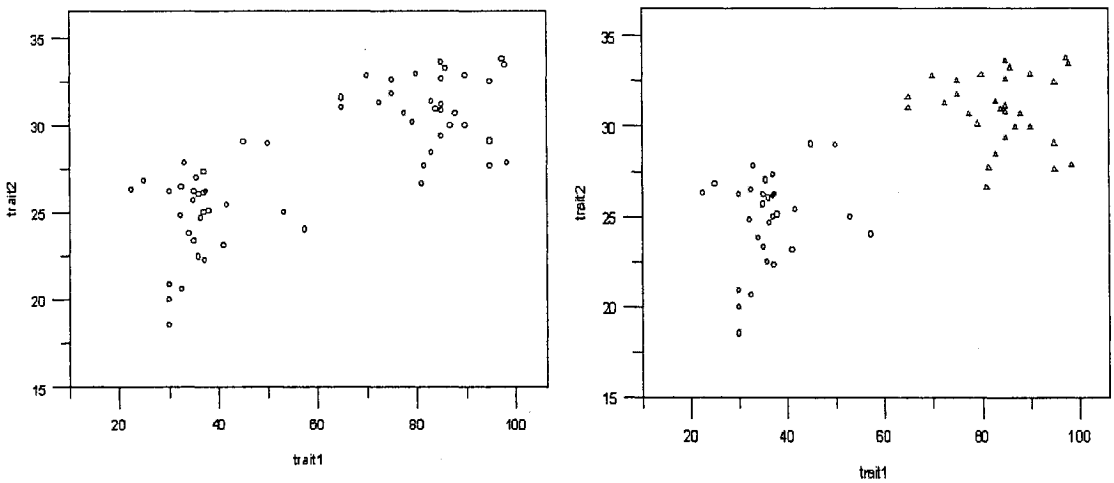


Fig. 1.3 Two traits appear to be correlated in cross-species scatter plot (left figure). However, if the phylogeny of the species shows they form two clades, with the round species being in one clade and the triangle species in another clade (right figure) then there is no correlation of the traits within either clade. Modified from Felsenstein, 2004.

Several statistical methods have been developed to take account of phylogenetic relatedness when doing cross-species comparisons, depending on discrete data or data of continuous variables (Harvey & Pagel, 1991). The key idea of these methods is that, while the characters of related species are not independent, the *changes* in characters along separate lineages are. For data of continuous variables, one of the most commonly used methods is the phylogenetic independent contrast (Felsenstein, 1985). The procedures to compute the contrasts are that 0) build a phylogenetic tree of the species, with correct topology and positive branch length; 1) for tree tips the contrast is the difference of traits of two adjacent tree leaves; 2) to compute traits of an upper node, assign to it the weighted average of its two lower nodes (or two tree leaves), the weights being proportional to the inverses of the variances of the two lower nodes; then compute the difference between two adjacent nodes or one node and a tree leave; 3) lengthen the

branch below the node by the product of the two lower node lengths divided by the sum of them. For original *n* species *n-1* contrasts will be extracted. When regressing the contrast for one trait on the contrast for another trait the regression line must pass through the origin of the two axes. The contrast method has been implemented in the phylogeny inference package Phylip (http://evolution.genetics.washington.edu/phylip.html; Felsenstein, 2004b). In Chapters 2 and 3 we will use both cross-species comparison and the contrast method to analyse rRNA composition and thermal adaptation in prokaryotes.

Another type of comparisons used in this thesis (Chapters 4 and 5) is to compare nucleotide composition, codon usage and amino acid usage between rice and *Arabidopsis* genomes. In those cases, we will compare the features of homologous genes and proteins between the two genomes, to see changes since the divergence of the two plant species from their common ancestor. The protein homologs are identified by comparing rice protein sequences against *Arabidopsis* protein sequences using BLAST (Altschul *et al.*, 1990) search and pairs of highest similarity score will be used for protein comparisons. The alignment of the homologous protein pairs will guard the alignment of protein coding genes between rice and *Arabidopsis*, which is then used for nucleotide composition and codon usage analysis.

1.6 Organization of the thesis

The thesis consists of six chapters. Following this introductory chapter, **Chapter 2** examines nucleotide composition and thermal adaptation in forty four bacterial and archaebacterial species. Consistent with previous studies, GC content of rRNA genes but not that of the whole genomes is found to correlate with temperature optimum in bacteria. We delineate the rRNA sequence into double stranded stem regions and single stranded loop regions and analysed GC contents of the two regions separately. While the GC content of the stems is strongly positively correlated with growth temperature, no correlation is found between the GC content of the loops and temperature. Furthermore, the loop GC content is also not correlated with genomic GC content. The nucleotide compositions of the loops are remarkably constant with adenine being most abundant regardless of genomic GC content or growth temperature. These results indicate that

thermal adaptation in bacteria mainly operates on the rRNA stems and there is a strong selective constraint on the loops.

Chapter 3 further extends the above analyses to all bacterial 16S rRNA genes. In addition to GC content we also consider the effect of temperature on gene length, as we reasoned that the longer the rRNA stems the more stable the molecule, that is rRNA length is also positively correlated with temperature. However, the relationship between GC content of rRNA and temperature was very weak for this much larger data set, in a sharp contrast with the results in Chapter 2 using a small data sample. The reason was that sample independence was violated in the statistical analyses. This is coherent in cross-species comparison, as the species are in a phylogenetically hierarchical structure and therefore traits of the species are not independent (Chapter 1.5). To solve this problem we used two methods to control the non-independence of the data samples. First, we compared nucleotide content and gene length in the same genus that have both mesophilic species and thermophilic species. This has found rRNA GC content is higher in the corresponding thermophilic species and the difference is significant. Second, we applied phylogenetic independent contrast methods (Felsenstein, 1985) to analyse the rRNA-temperature relationship in archaeal species. This has confirmed the previous findings of a strong positive correlation between the rRNA GC content and temperature and also found evidence of a weaker positive relationship between rRNA length and temperature. This chapter also includes a comparison of GC content and gene length of 18S rRNA genes between warm-blooded and cold-blooded vertebrates.

Chapter 4 examines the effect of nucleotide bias on protein evolution in two angiosperm plants, rice and *Arabidopsis*. Homologous proteins and genes between the two genomes are identified by BLAST search and used in the two genome comparisons. The rice genome has certain peculiar features that the *Arabidopsis* genome lacks. Rice genes can be separated into two GC content classes: high GC and low GC genes, and the rice genes have a gradient in GC content: the GC content at the 5' part is higher than that at the 3' part. These features allow us to analyse the GC content effects on different scales: between the two genomes (rice and Arabidopsis), within a genome (high GC and low GC rice genes) and within a gene (5' and 3'parts of a rice sequence). This study indicates that rice genes that have increased GC content show a corresponding,

predictable change in the amino acid compositions of the encoded proteins, and the biased pattern of protein evolution is the consequence, rather than the cause, of the corresponding change in nucleotide content.

Chapter 5 focuses on the effect of nucleotide bias on codon usage in the rice genome, with a comparison to the codon usage of homologous *Arabidopsis* genes. These analyses found synonymous codon usage is primarily dictated by the GC content of the gene, and translation selection plays a negligible role in the rice genome.

The last chapter, **Chapter 6,** contains a general discussion of the results and suggests future directions in the research of nucleotide bias and genome evolution.

Chapter 2

Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes*

2.1 Abstract

Previous studies have shown that the guanine plus cytosine (G+C) content of ribosomal RNAs (rRNA) is highly correlated with bacterial growth temperatures. This correlation is strongest in the double-stranded stem regions of the rRNA, a fact that can be explained by selection for increased structural stability at high growth temperatures. In this study, we examined the single stranded regions of 16S rRNAs. We reasoned that, since these regions of the molecule are subject to less structural constraint than the stem regions, their nucleotide content might simply reflect the overall nucleotide content of the genome. Contrary to this expectation, however, we found that all of the single stranded regions are characterized by very high adenine and relatively low cytosine contents. Moreover, the nucleotide content of these single-stranded regions is surprisingly constant between species, despite dramatic differences in optimal growth temperatures, and despite large differences in the overall genomic G+C content. This provides compelling evidence for strong stabilizing selection acting on 16S rRNA single stranded regions. We found that selection favors purines (A+G), and especially adenine, in the single stranded regions of these ribosomal RNAs.

---

* Adapted from Wang H-C. and D.A. Hickey (2002) *Nucleic Acids Research* 30: 2501-2507.

2.2 Introduction

The thermal stability of double-stranded DNA is dependent on the nucleotide content of the molecule. Specifically, DNA molecules that are rich in guanine and cytosine are more thermostable than those with an excess of adenine and thymine. This is because G:C base pairs have an additional hydrogen bond compared to A:T pairs; consequently, it has been proposed that the presence of extra G:C pairs should help stabilize DNA and RNA secondary structures at elevated temperatures (Wada & Suyama, 1986). This has led to the further prediction that there should be a correlation between the G+C content of genomes and environmental temperature. Recent studies of prokaryotic genomes, however, failed to demonstrate any correlation between the overall G+C content of the genome and optimal growth temperature (Dalgaard & Garrett, 1993; Galtier & Lobry, 1997; Hurst & Merchant, 2001). Nevertheless, there is a very strong relationship between the G+C content of structural RNAs (including small subunit ribosomal RNA, large subunit ribosomal RNA, transfer RNA and 5S RNA) and bacterial growth temperature (Dalgaard & Garrett, 1993; Galtier & Lobry, 1997). In the case of the ribosomal RNAs, the elevated G+C content was concentrated primarily in the double-stranded stem regions of the molecule, and it was largely independent of the average G+C content of the bacterial genome (Galtier & Lobry, 1997; Hurst & Merchant, 2001). These results indicate that, while thermal adaptation does not affect the overall nucleotide content of the genome, it has a very significant effect on the composition of the double-stranded regions of structural RNAs.

In this study, we focused on the single stranded regions of the 16S ribosomal RNA genes. We calculated the nucleotide compositions of these regions in both mesophilic and thermophilic eubacteria and archaea. We were especially interested in the covariation with both optimal growth temperature and the average nucleotide content of the genome. Our expectation was that, in contrast to the double-stranded stems, these unpaired regions of the rRNA molecule would be affected by the same mutational pressures that determine the overall nucleotide content of the genome. As a control, we also scored the nucleotide content of the genes encoding ribosomal proteins in these

genomes. This study was made possible by the recent availability of complete genomic sequences for many bacterial species.

2.3 Materials and Methods

We assembled a database that contained the nucleotide contents and optimal growth temperatures for 44 prokaryotic species. Nucleotide contents of the total 16S rRNAs, along with their component secondary structures (stems, loops, bulge loops and internal loops), were computed from the European Small Subunit RNA Database (http://rrna.uia.ac.be; Van de Peer et al. 2000). Ambiguous nucleotides, which account for about 4.8% of an average sequence length (1520 nucleotides), were ignored in the calculation. An RNA stem is defined as a right handed double helix of base pairs; a hairpin loop is a loop of unpaired nucleotides at the termini of stems; a bulge loop is formed by unpaired nucleotides in one strand of a double-stranded region, where the other strand has contiguous base pairing; and an internal loop contains several unpaired nucleotides in both strands of a double-stranded region (De Rijk, 1995). Stems, bulge loops and internal loops are each identified in this database. The remainder of the molecule consists of hairpin loops, multiple branched loops, pseudoknot loops and dangling ends (5' terminal and 3' termini). These latter regions, together with the bulge loops and internal loops, are collectively called single stranded regions in this study.

In all, we have data from 28 mesophilic species (optimal growth temperature less than 45°C) and 16 thermophilic species (growth temperature equal to or greater than 45°C). The entire genomes of 31 of these species (21 eubacteria and 10 archaea) have been completely sequenced and these genomic sequences were retrieved from GenBank. For these species, in addition to obtaining data on the 16S rRNA sequences, we computed nucleotide compositions for the entire genome and for the genes encoding ribosomal proteins. Growth temperature data on the thermophiles having complete genomic sequences available were taken from the original papers describing the genomic sequence. Growth temperature data for the remaining species were mainly obtained from the data sets described in (Galtier & Lobry, 1997; Hurst & Merchant, 2001) and DSMZ German Collection of Microorganisms and Cell Cultures

(http://www.dsmz.de/species/strains.htm). Additional data may be viewed at http://www.bact.wisc.edu/microtextbook/NutritionGrowth/Temperature.html.

Statistical analyses were performed using the statistics package SYSTAT version 10 (SPSS Science, 2000). For the correlation analyses, the Pearson correlation coefficient r was used to evaluate the strength of correlation. A "strong" correlation is defined as an absolute value of r greater than 0.9; a value between 0.5 and 0.9 represents a "moderate" correlation; and a value of less than 0.5 is defined as a "weak" or no correlation (Milton, 1992). A negative value of r means that the correlation is negative.

2.4 Results

2.4.1 Average nucleotide composition in mesophiles and thermophiles

We first compared the average nucleotide composition of the entire genome with the optimal growth temperature for each species. Specifically, we wished to test for a correlation between the G+C content of the entire genome and optimal growth temperature. The results are shown in Table 2.1. As can be seen in the Table, although there is a wide variation in G+C content within both the mesophiles and the thermophiles, there is no obvious difference between the two groups. In fact, we found no significant difference in the average genomic G+C content between these two sets of species (Mann-Whitney U-test, p=0.811). Next, we compared the average nucleotide content of the entire genome to that of the 16S ribosomal RNA sequences (see Table 2.2). In this case, we calculated the frequency of each of the four nucleotides separately. Again, we found that mesophiles and thermophiles have very similar base compositions when we score the entire genome (see Table 2.2A). For example, the mean total genomic adenine content of the 22 mesophiles is 27.2±1.5%; while the mean genomic adenine content of the 9 thermophiles is very similar at 27.8±1.3%, and the values are not significantly different (Mann-Whitney U-test, p=0.794). Likewise, there is no significant difference in the average frequency of the other three bases between the genomes of the mesophiles and the thermophiles. The result is quite different, however, when we compare the nucleotide compositions of 16S rRNA genes. In this case, there is a clear, and highly significant

Table 2.1 GC content and optimal growth temperature (T_{opt}, °C) of completely sequenced genomes used in this study.

Organism	G+C (%)	T_{opt}	References*
A) Mesophile			
Bacillus subtilis	43.5	38.8	1
Borrelia burgdorferi	28.6	37.0	2
Campylobacter jejuni	30.5	37.0	2
Caulobacter crescentus	67.2	22.5	1
Chlamydophila pneumoniae	40.6	37.0	3
Chlamydia trachomatis	41.3	37.0	3
Deinococcus radiodurans	66.6	37.0	1
Escherichia coli	50.8	37.0	1
Haemophilus influenzae	38.1	36.0	1
Halobacterium sp.	67.9	37.0	2
Helicobacter pylori	38.9	37.0	2
Mycoplasma genitalium	31.7	37.0	3
Mycoplasma pneumoniae	40.0	37.0	1
Mycobacterium tuberculosis	65.6	37.0	1
Neisseria meningitides	51.5	36.0	1
Pasteurella multocida	40.4	37.0	1
Pseudomonas aeruginosa	66.6	37.0	1
Rickettsia prowazekii	29.0	37.0	3
Streptococcus pyogenes	38.5	37.0	1
Treponema pallidum	52.8	37.0	4
Vibrio cholerae	47.5	37.0	4
Ureaplasma urealyticum	25.5	37.0	1
Mean ± Standard Error	45.6±2.9	36.3±0.7	
B) Thermophile			
Aeropyrum pernix	56.3	90.0	5
Aquifex aeolicus	43.5	95.0	6
Archaeoglobus fulgidus	48.6	83.0	7
Methanobacterium thermoautotrophicum	49.5	65.0	8
Methanococcus jannaschii	31.4	85.0	9
Pyrococcus horikoshii	41.9	95.0	10
Sulfolobus solfataricus	35.8	80.0	11
Thermoplasma acidophilum	46.0	59.0	12
Thermotoga maritime	46.2	80.0	13
Mean ± Standard Error	44.4±2.5	81.3±4.1	

* 1, Galtier & Lobry, 1997. 2. DSMZ, http://www.dsmz.de/species/strains.htm. 3. The optimal growth temperature of this species was not available from the temperature data we collected. However, since it is an obligate intracellular pathogen of humans or mammals, its growth temperature was assumed to be 37°C in this study. 4. http://www.bact.wisc.edu/microtextbook/NutritionGrowth/Temperature.html. 5. Kawarabayasi et al., 1999. 6. Deckert et al., 1998. 7. Klenk et al., 1997. 8. Smith et al., 1997. 9. Bult et al., 1996. 10. Kawarabayasi et al., 1998. 11. She et al. 2001; 12. Ruepp et al., 2000. 13. Nelson et al., 1999.

difference between mesophiles and thermophiles (see Table 2.2B). For example, the average adenine content of the 16S rRNA sequences among the mesophiles is 26.1±0.4%, while the mean adenine content among the thermophiles is only 20.3±0.4%, and this difference is statistically significant (Mann-Whitney *U*-test, $p<0.001$). A similar difference between mesophiles and thermophiles is seen for the other three bases. Overall, the 16S rRNA sequences of the thermophiles are relatively G+C rich, as expected, and we note that, in mesophiles, the purines (G and A) are two most abundant bases.

Table 2.2 Average nucleotide composition (mean% ± standard error) of whole genomes and 16S rRNA genes of mesophiles and thermophiles.

A) Nucleotide composition of entire genome, for the 31 genomes listed in Table 1.

Nucleotide	Mesophile	Thermophile	Significance[a]
A	27.2±1.47	27.8±1.26	NS*
C	22.8±1.48	22.1±1.26	NS
G	22.8±1.46	22.2±1.22	NS
T	27.2±1.47	27.8±1.22	NS

B) Nucleotide composition of 16S rRNA genes

Nucleotide	Mesophile	Thermophile	Significance[a]
A	26.1±0.38	20.3±0.41	$p < 0.001$
C	21.6±0.31	28.8±0.49	$p < 0.001$
G	30.5±0.38	36.0±0.47	$p < 0.001$
T	21.8±0.31	14.9±0.55	$p < 0.001$

[a] Probability of Mann-Whitney *U*-test of respective nucleotide composition in the two sets of mesophiles and thermophiles. * NS, not significant ($p>0.5$).

Since our main focus was on the unpaired regions of the 16S rRNA sequence, we subdivided the rRNA molecule according to its secondary structure, and the base compositions of these structural components are summarized in Table 2.3. In the double-stranded stem regions (Table 2.3A), although both mesophiles and thermophiles are rich in G and C, the frequency of these two bases is significantly higher in thermophiles (Mann-Whitney *U*-test, $p<0.001$). These results are consistent with the hypothesis that G:C base pairs are selected because of their role in stabilizing the stem regions, and that

Table 2.3 Average nucleotide composition (mean% ± standard error) of structural components of 16S rRNA genes of 44 mesophiles and thermophiles.

A) Nucleotide composition of the stems; the average cumulative length of stems in a 16S rRNA is 856 nucleotides*.

Nucleotide	Mesophile	Thermophile	Significance[a]
A	15.7±0.46	7.5±0.59	$p < 0.001$
C	26.7±0.42	37.0±0.74	$p < 0.001$
G	34.9±0.45	42.8±0.58	$p < 0.001$
T	22.7±0.41	12.7±0.73	$p < 0.001$

B) Nucleotide composition of the loops (including hairpin loops, multiple branched loops, psudoknot loops and dangling ends); the average cumulative length of loops in a 16S rRNA is 427 nucleotides*.

Nucleotide	Mesophile	Thermophile	Significance
A	37.6±0.35	37.8±0.34	NS**
C	16.5±0.24	17.8±0.21	$p = 0.001$
G	23.9±0.28	25.1±0.37	$p = 0.012$
T	21.9±0.26	19.2±0.43	$p < 0.001$

C) Nucleotide composition of the internal loops and bulge loops; the average cumulative length of internal/bulge loops in a 16S rRNA is 163 nucleotides*.

Nucleotide	Mesophile	Thermophile	Significance
A	48.2±0.46	45.5±0.42	$p = 0.001$
C	9.2±0.30	11.3±0.46	$p = 0.001$
G	25.5±0.59	27.1±0.55	NS
T	17.0±0.56	16.0±0.40	NS

[a] Probability of Mann-Whitney *U*-test of respective nucleotide composition in the two sets of mesophiles and thermophiles.

* Ambiguous nucleotides were excluded from the calculation.

** NS, not significant ($p>0.05$).

this selection is especially strong among the thermophiles. When we look at the unpaired regions of the 16S rRNA, i.e., the single-stranded loops, bulges and dangling ends, a very different picture emerges (see Table 2.3B and Table 2.3C). We note two trends: first, these segments of the rRNA sequence are not particularly rich in G+C, and secondly, there is not a large difference in any of the four base frequencies between the thermophiles and the mesophiles. For instance, whereas there is a two-fold difference in adenine content of the stem regions between the mesophiles and thermophiles (Table

2.3A), there is no significant difference between mesophiles and thermophiles as regards their adenine content in the loops and dangling ends (Mann-Whitney *U*-test, p=0.893, see Table 3B). Likewise, although there are differences for the other three bases, the magnitude of these differences is relatively small. What is most striking about the data on the single stranded regions are the high levels of adenine and of purines in general (A and G) in both the mesophiles and the thermophiles.

In order to explore these patterns further, we examined the distribution of nucleotide frequencies in the 16S rRNA sequences from the individual species. We were especially interested in the correlations between these nucleotide frequencies and (i) the optimal growth temperature of the species and (ii) the average nucleotide content of the entire genome.

2.4.2 16S rRNA stems and loops are affected very differently by growth temperature

We plotted the relationship between the G+C content of the 16S rRNA genes and the optimal growth temperatures of the 44 species. For this plot, we separated the data into two sets: one set for the double-stranded stem regions, and the other for the single stranded regions. The results are shown in Figure 2.1. As can be seen in the Figure, the G+C content of the stem regions rises rapidly with increasing temperature (slope of regression line = 0.34, p<0.001). A much weaker trend, however, is seen for the dependence of the G+C content of the single stranded regions on growth temperature (slope of regression line = 0.06). From this result, it is clear that the selection for an increased G+C content of 16S rRNA sequences at higher temperatures is operating almost entirely within the stem regions. It is interesting to note that the two species with moderate thermophilicity (*T. acidophilum* , with an optimum growth temperature of 57°C and *M. thermoautotrophicum* , with an optimal temperature of 65°C) are also intermediate with respect to the nucleotide content of their rRNA stem regions.

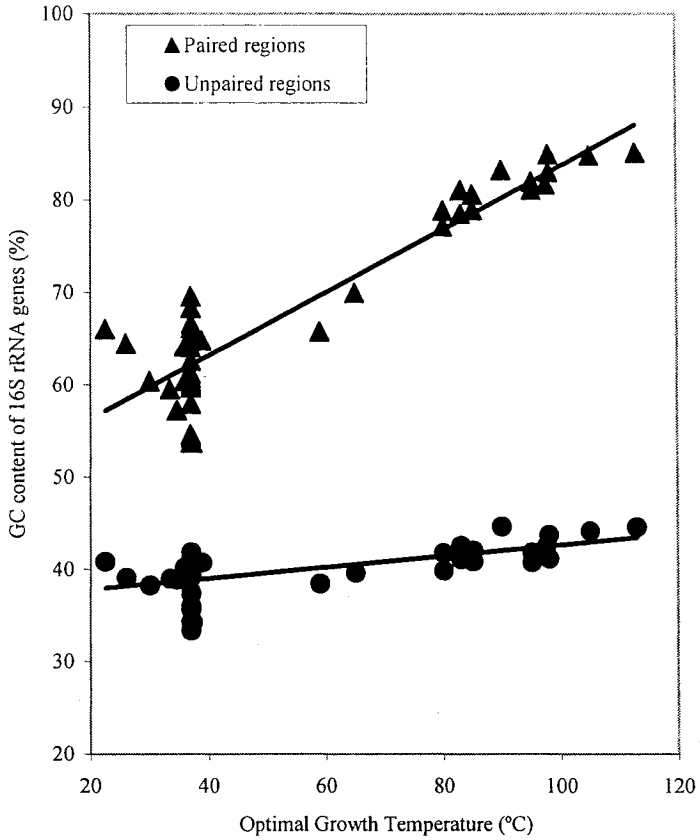


Fig. 2.1 G+C contents of 16S rRNA paired regions (stems) and unpaired regions (single strand regions) plotted against optimal growth temperature (°C). The slope of the regression line for $G+C_{paired\ regions}$ vs. temperature is 0.34. The slope of the line for $G+C_{unpaired\ regions}$ vs. temperature is only 0.06.

In order to get more detailed information on the patterns of nucleotide distribution in the single stranded regions, we plotted the frequencies for each of the four bases separately (Fig. 2.2). In this Figure, we can observe the slight increase in the frequencies of G and C with increasing growth temperature, and the concomitant decreases in the frequencies of A and T. What is much more marked than these slight changes related to growth temperature, however, are the very large and relatively constant differences in the frequencies of the four bases. Adenine is at a uniformly high frequency, followed by guanine, whereas the two pyrimidines, cytosine and thymine are at low frequencies at all growth temperatures.

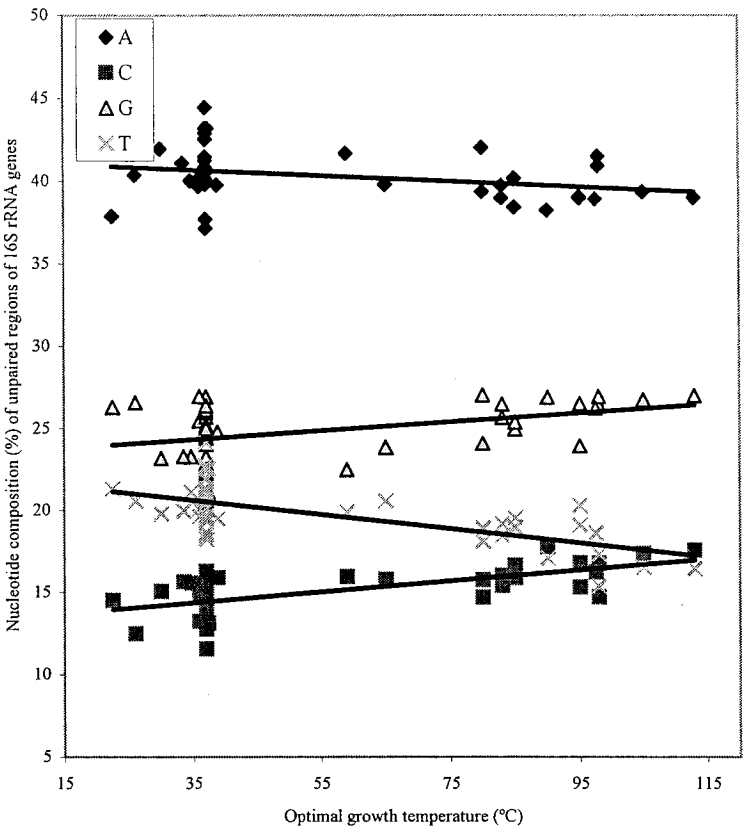


Fig.2.2 Individual nucleotide composition of unpaired regions of 16S rRNA genes plotted against optimal growth temperature (°C).

In summary, these data show that nucleotide contents in single stranded regions of the 16S rRNA are relatively constant between the mesophiles and thermophiles, and they show that adenine is the most abundant nucleotide in both groups. In contrast to this weak relationship between the nucleotide content of the single-stranded regions and optimal growth temperature, there is a very strong temperature dependence in the double-stranded stem regions (compare Table 2.4A and Table 2.4B). In the case of the paired regions, the G and C contents are strongly, and positively, correlated with growth temperature, whereas both A and T show a strong negative correlation with growth temperature in these regions (Table 2.4A).

Table 2.4 Correlation and regression analysis of nucleotide composition of 16S rRNA and optimal growth temperature.

	r[a]	Slope[b]	p value on slope[c]
A) Paired Regions			
A	-0.891	-0.153	$p < 0.001$
C	0.931	0.192	$p < 0.001$
G	0.889	0.147	$p < 0.001$
T	-0.928	-0.187	$p < 0.001$
B) Unpaired Regions			
A	-0.297	-0.017	0.051
C	0.643	0.033	$p < 0.001$
G	0.412	0.027	0.005
T	-0.657	-0.043	$p < 0.001$

[a] Pearson correlation coefficient.
[b] The slope of the linear regression line.
[c] The associated probability associated with the null hypothesis (regression line slope=0).

2.4.3 The relationship between the nucleotide content of the 16S rRNA and the nucleotide content of the whole genome

It is already known that the nucleotide content of the stem regions of the rRNAs are distinct from the background genomic levels (Galtier & Lobry, 1997). Our working hypothesis was that the loop regions, which are not subject to selection for elevated G+C levels, might show a pattern of nucleotide composition that resembled the background genomic levels. Since we now have complete genomic sequences for many species, we were able to test this hypothesis directly. In Figure 2.3, we show the relationship between the G+C content of the 16S rRNA and the average G+C content of the corresponding whole genome. For this analysis the data were separated into stem and loop regions, and into thermophiles and mesophiles. Although there is a significant positive relationship between G+C content of the stem regions and total genomic G+C content among the mesophiles (slope of the regression line = 0.29, $p<0.001$), there is no such relationship for the stem regions of the thermophiles (slope of the regression line = 0.149, $p = 0.532$). To our surprise, we found that the relationship between the nucleotide content of the single-stranded regions and that of the whole genome was extremely weak.

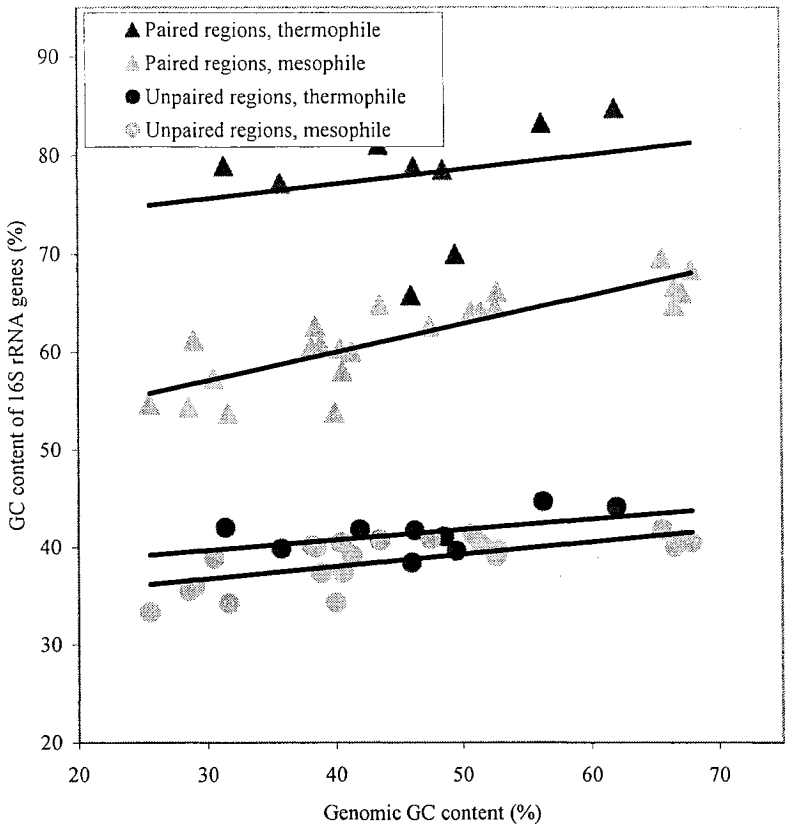


Fig. 2.3 G+C content of 16S rRNA paired and unpaired regions plotted against the average G+C content of the whole genome. Data for mesophiles and thermophiles are shown separately.

Whereas Figure 2.2 shows that the nucleotide composition of the single stranded regions is not very sensitive to changes in optimal growth temperatures, Figure 2.3 shows that these regions are equally immune to the effects of large changes in the nucleotide content of the whole genome. To highlight this relative constancy of nucleotide content in very different genomic backgrounds, we compared the patterns of nucleotide content of the single-stranded regions with that of the ribosomal protein genes from the same species. The results are shown in Figure 2.4 and in Table 2.5. Unlike the single stranded regions of the rRNA, the ribosomal protein coding genes show a very strong effect of the genomic background among both the mesophiles and the thermophiles (the slopes of the regression lines are 0.787 and 0.805, respectively). This difference between the rRNA and the ribosomal protein genes reinforces the impression that there is a positive selection

maintaining the relatively constant nucleotide levels in the single stranded regions of the rRNAs.

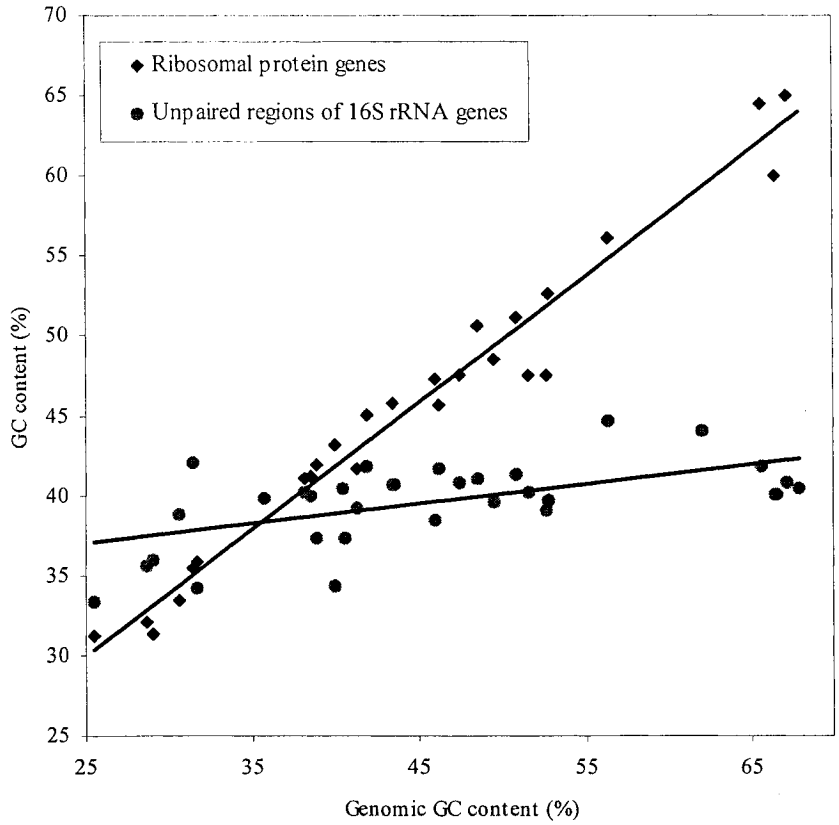


Fig. 2.4 The relationship between the average genomic G+C content and the G+C content of (i) ribosomal protein genes and (ii) 16S rRNA unpaired regions.

Table 2.5 The relationship between the overall G+C content of the genome and the G+C content of A) 16S rRNA unpaired regions and B) the ribosomal protein gene coding sequences.

	r[a]	Slope[b]	p value on slope[c]
A) 16S rRNA unpaired regions			
Mesophile	0.689	0.126	$p < 0.001$
Thermophile	0.493	0.107	0.147
B) Ribosomal protein coding sequences			
Mesophile	0.985	0.787	$p < 0.001$
Thermophile	0.981	0.805	$p < 0.001$

[a] Pearson correlation coefficient.
[b] The slope of the linear regression line.
[c] The associated probability associated with the null hypothesis (regression line slope=0).

2.5 Discussion

The main finding of this study is that the nucleotide content of the single stranded regions of 16S rRNA (hairpin loops, multiple branched loops, pseudoknot loops, internal loops, bulge loops and dangling ends) are remarkably constant between thermophiles and mesophiles, and also among genomes that have very different average nucleotide frequencies. Essentially, these regions are characterized by a uniformly low G+C content in all of the species studied. At first glance, this seems to contradict the previous studies, which showed that the G+C contents of ribosomal RNA genes are positively correlated with bacterial growth temperature (Dalgaard & Garrett, 1993; Galtier & Lobry, 1997). This seeming paradox can easily be resolved, however, by the realization that the increased G+C content of the thermophiles is concentrated almost entirely within the double-stranded stem regions of the molecule.

Since the sequences within the loop regions do not contribute to the formation of the secondary structure of the rRNA molecule, it is not surprising that they are not affected by temperature-based selection in the same way as the stem regions. One might expect, however, that they would reflect the nucleotide biases that affect the genome as a whole. It has now been well established that such biases in nucleotide composition can have a major effect on the coding capacity of most genes within the genome, causing predicable changes in the amino acid composition of the encoded proteins (Singer & Hickey, 2000; Kreil & Ouzounis, 2001). Given these findings, we reasoned that those regions of the rRNA, which were not critical for forming the secondary structure, might respond to nucleotide pressures that affect the genome as a whole. Based on our results presented here, this is clearly not the case. As shown in Figure 2.4, even the highly conserved ribosomal protein-coding sequences are affected by the overall nucleotide content of the genome, but this is not the case for the single stranded regions of the ribosomal RNA. This indicates that there are very strong selective constraints acting on these single stranded regions to maintain their constant nucleotide frequencies.

An examination of the frequencies of the individual nucleotides gives some clues about the nature of these selective forces. In single stranded regions, we found that adenine is the most abundant nucleotide base followed by guanine, thymine and cytosine

(A>> G ≥ T ≥ C). On the other hand the frequency of adenine is lowest in the stem regions (Table 2.3A). One possible explanation might be that selection for high G and C levels in the stem regions has resulted in a local depletion of these two nucleotides in the free nucleotide pool, thus leaving an excess of A and T. We can exclude this possibility, however, by noting that the free nucleotide pool for each species varies over a wide range, as indicated by the variation in the nucleotide content of the whole genome. A more probable explanation of high level of A (and also G) in the unpaired regions of rRNA in all species is that it plays an essential role in maintaining the secondary and tertiary structure of the molecule.

Ribosomal RNA may contain many non-canonical base pairs, other than Watson-Crick pairs (http://prion.bchs.uh.edu/bp_type/bp_structure.html). Purines are especially good at forming base triples and non-canonical pairs such as sheared GA, GA imino, Hoogsteen, reverse Hoogsteen and wobble pairs (Nagaswamy et al., 2002). The high amount of A and G in the single stranded regions may be involved in tertiary interactions between these regions (Gutell et al., 2000). The detailed structures of 16S rRNA of *Thermus thremophilus* and 23S rRNA of *Haloarcula marismortui* are now known (Wimberly et al., 2000; Ban et al., 2000). It will be interesting to examine these structures and to make an inventory of tertiary interactions between nucleotides of single stranded regions.

We wondered if the characteristically high G+C contents of ribosomal RNA genes among the thermophiles could be used as a diagnostic test of thermophily. To test this idea, we calculated and sorted G+C contents of all sequence entries deposited in the small subunit rRNA database of Ribosomal Database Project (http://rdp.cme.msu.edu/). Species with the high rRNA G+C contents are indeed thermophiles. The two species having the highest G+C contents in their rRNAs (68.99% in *Pyrolobus fumarius* and 68.81% in *Pyrodictium occultum*) were found to be hyperthermophiles (*Pyrolobus fumarius* can live at temperatures up to 113°C; *Pyrodictium occultum* grows at 85-105°C). A number of recent reports (Rivas & Eddy, 2000; Carter et al., 2001; Klein, Misulovin & Eddy, 2002) have proposed that structural RNAs could be identified as high G+C islands in a low G+C genomic background. We have confirmed that the 16S rRNA sequences correspond to such islands in the genomes of *M. jannaschii*, *M. genitalium* and *B. burgdorferi* (all of

which have A+T rich genomes). For example, Figure 2.5 shows a G+C plot of *M. jannaschii* genome, calculated in a nonoverlapping window of 5000 nucleotides. The genomic G+C content is basically consistent, with 90% confidence intervals fall between 28.52% and 34.44%. The highest peak in the graph was subsequently identified to contain 23S rRNA, 16S rRNA, and arginine tRNA genes, all of which are structural RNA genes.

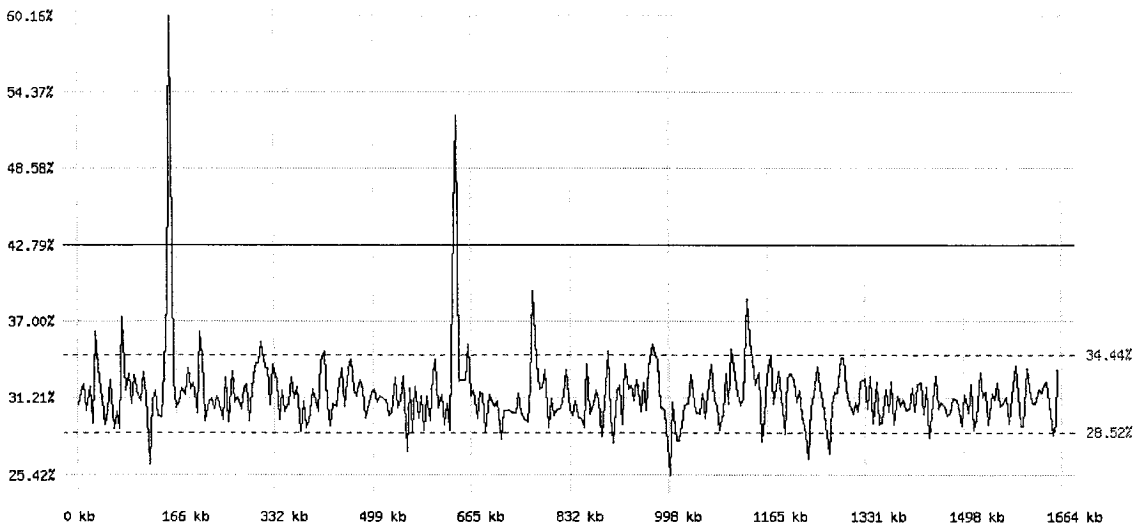


Fig. 2.5 A G+C plot of *M. jannaschii* genome (from http://www.tigr.org/tigr-scripts/CMR2/GCDisplay.spl?asmbl_id=259). The genome is represented on the X-axis from left to right. The Y-axis is the average GC content for a non-overlapping window 5000 nucleotides. The black center line is the median GC content of this genome. The two dashed lines represent the 5% lower limit and the 95% upper limit of the average GC content. The highest peak on the left contains rRNA genes and tRNA genes.

Our finding of very stable nucleotide contents among 16S rRNA sequences, despite large variations in the nucleotide content of the entire genome, is intriguing in light of an earlier report (Hasegawa & Hashimoto, 1993) that the small subunit ribosomal RNA sequences are subject to compositional biases that can affect phylogenetic inference based on these sequences. A re-analysis of three protist sequences (*Giardia lamblia*, *Vairimorpha necatrix* and *Entamoeba histolytica*) that were used in that earlier study showed that there are, indeed, very large differences in G+C content among these ribosomal RNAs. For instance, the rRNA of *G. lamblia* contains 74.4% G+C, which is equivalent to that seen among the bacterial thermophiles. In contrast, the G+C content of

the 16S rRNA in *E. histolytica* is only 38.3% and that in *V. necatrix* is 37.4%. This suggests that the strong constraint on nucleotide content that is seen among the prokaryotes may not apply to eukaryotes.

In summary, we have used the recently-available large genomic datasets to confirm earlier reports that the double stranded regions of rRNA are subject to intense selection for increased G+C content, especially in organisms with high optimal growth temperatures, and even in an A+T-rich genomic background. More important, we have found that the single-stranded regions of the rRNA are subject to equally intense selection to maintain a very constant nucleotide composition, in the face of large variations in both optimal growth temperature and nucleotide composition of the whole genome. The fact that this pattern in seen in both eubacteria and archaea provides further evidence for the action of selection, as opposed to an accident of phylogenetic history. This selective force strongly favors purines, and especially adenine, in the unpaired regions of the rRNA. The biochemical basis for this selective preference remains to be elucidated, although it is tempting to speculate that these nucleotides are critical for the maintenance of higher order rRNA structure.

Chapter 3

Thermal adaptation of ribosomal RNA genes

3.1 Abstract

The main finding reported in Chapter 2 was that there is strong conservation of the nucleotide composition and especially abundant adenines in the unpaired regions of 16S ribosomal RNAs. In this chapter, we carried out a comprehensive survey of rRNA sequences from a wide range of lineages in order to understand the general patterns of thermal adaptation in these genes. We compared the nucleotide content and sequence length of small subunit rRNAs between mesophilic, moderately thermophilic and hyperthermophilic bacteria and archaea, and we also compared these genes between warm-blooded vertebrates and cold-blooded vertebrates. We reexamined the relationship between optimal growth temperature and the nucleotide contents of 16S rRNA, while correcting for the phylogenetic relatedness of the RNA sequences. Specifically, we compared the GC content and length of the rRNA of mesophilic and thermophilic species within the same genus and also used the method of independent contrasts to avoid phylogenetic dependence when doing cross-species comparisons. We confirmed the previous findings of a strong positive correlation between the rRNA GC content and environmental growth temperature in prokaryotes. We also found evidence of a weaker positive relationship between rRNA length and temperature. Therefore, we can conclude that the positive relationship between the nucleotide content and length of ribosomal RNAs and environmental growth temperature of bacteria and archaea is not due to phylogenetic history, but reflects a repeated selective response to elevated environmental temperature. Warm-blooded vertebrates also have higher rRNA GC content and length than cold-blooded vertebrates, but this is not concentrated in the paired regions of the molecule; this argues against thermal adaptation as an explanation for the differences in the vertebrate sequences.

3.2 Introduction

Structural RNAs such as rRNAs and transfer RNAs are less subject to mutation-based nucleotide bias than other genes. For instance, the overall genomic GC content of 529 bacterial and archaeal species varies between 23% and 77%, while the range of the corresponding 16S rRNA GC content is only 45% to 69% (data based on Galtier & Lobry, 1997). Previous studies have shown that the GC content of 16S ribosomal RNAs (and especially of the stem regions) is highly correlated with optimal growth temperature in bacteria (Dalgaard & Garrett, 1993; Galtier & Lobry, 1997; Hurst & Merchant, 2001; Wang & Hickey, 2002; Nakashima et al., 2003). However, when we applied the approach to 16S rRNAs of 1573 bacterial species (described in 3.3.1 and 3.3.2) we found the GC content and temperature correlation was very low for this large data set, even for the stems (Fig. 3.1). Several factors may account for this low correlation. First, the data set is particularly unbalanced. There are large clusters of points (species) located at the same growth temperatures (at 28°C, 30°C and 37°C), but with large variation in their GC

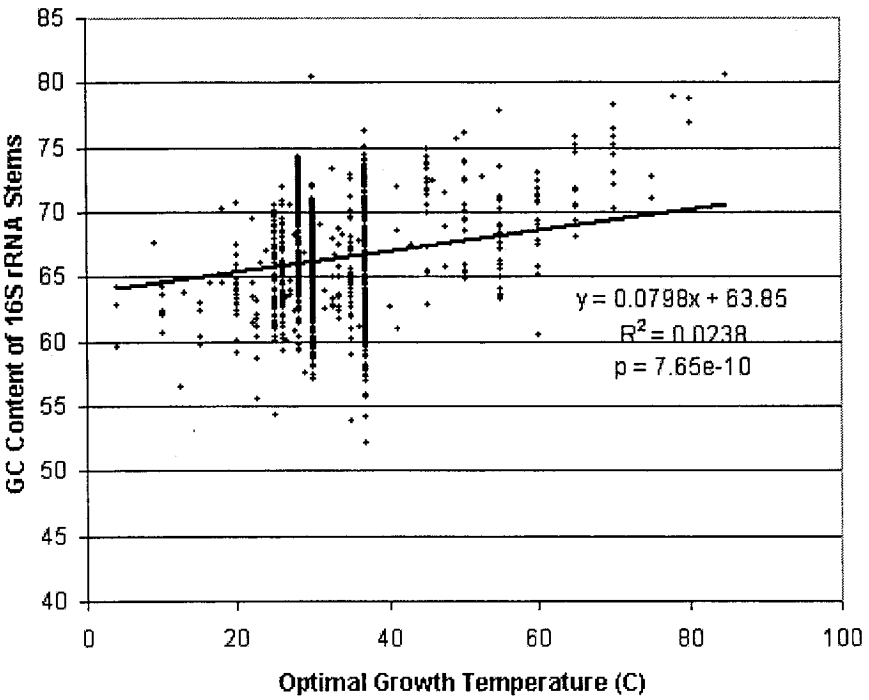


Fig. 3.1 Correlation of GC content of 16S rRNA stem regions and optimal growth temperature in 1573 bacterial species. A positive correlation is only evident in the higher temperature species.

contents. For instance 425 species have a growth temperature of 37 °C and the GC content of the rRNA stems of the species varies from 52 to 76%. Because these are the major elements of the data the relationship of GC stem and temperature is almost reduced to null for all species (Fig. 3.1). The relationship was much stronger when the 1109 species that have a temperature of 28, 30 and 37 °C were removed. The regression R^2 for the stem GC and temperature for the remaining 464 species is dramatically increased to 0.31 from the original 0.0259. But it is still much weaker than the correlation seen in Chapter 2 (R^2 = 0.84; Fig. 2.1). This is because of a second reason, and the most important reason, that the data sample is not independent. As described in Chapter 1.5, the species are part of a hierarchically structured phylogeny and, consequently, they cannot be regarded as independent samples from the same distribution (Felsenstein, 1985; Harvey and Pagel, 1991). Therefore, in this chapter, we tried to minimize the phylogenetic influence on the cross-species comparison by using phylogeny-based comparative methods and reexamined the relationship between growth temperature and the nucleotide content of 16S rRNA.

We also studied the effect of bacterial growth temperature on the length of its 16S rRNA gene. This is based on the following considerations. First, it is known that sequence length and GC content in protein-coding genes are correlated (Duret, Mouciroud & Gautier, 1995; Oliver & Marin, 1996; Carels & Bernardi, 2000; Xia et al., 2003). Therefore, statistically the 16S rRNA gene length may confound the effect of temperature on GC content. Second, spontaneous deamination of cytosine (C) occurs frequently at high temperature, which leads to a cytosine to urosine (C to U) change. If C is originally paired with a guanine (G) in an rRNA stem, after the C to U substitution, the C:G pair will be replaced by a U:G pair. This U:G pair is even weaker than an A:T pair, leading to the instability of the rRNA. Therefore, one may postulate that an rRNA will require not only high GC content in its stem regions but also longer stems to sustain higher temperature. We tested the relationship between rRNA length and the optimal temperature with and without phylogenetic relatedness considered.

Furthermore, we extended the analyses to vertebrate 18S rRNA genes and compared the GC content and gene length between the cold-blooded and warm-blooded vertebrates.

3.3 Methods

3.3.1 Sequence data

The rRNA sequence data were downloaded from the ssu rRNA database http://oberon.fvms.ugent.be:8080/rRNA/ssu/index.html (Wuyts et al., 2002). For prokaryotes, there are 10566 eubacterial and 590 archaeal 16S rRNA sequence entries (data up to July 2003). For each rRNA sequence, the GC content of the whole molecule as well as that for the stems and loops, respectively, were calculated (see Chapter 2.3). The rRNA length and the cumulative lengths of the stems and loops, respectively, were also calculated. Since many species have multiple entries of 16S rRNA sequences in the database, the rRNA data were cleaned by retaining the first entry when there are multiple entries for the same species, which left rRNA data of 4598 species retained. If the rRNA data of same species were cleaned by averaging values of same species and the mean values for each species were used, the subsequent analyses gave very similar results as the above method to clean multiple entries of same species.

For vertebrates, 18S rRNA sequences of 84 species in the rRNA database were collected, of which 38 species are warm-blooded animals, including 34 birds and 4 mammals, and 36 species are cold-blooded animals, including 23 fishes, 18 amphibians and 5 reptiles. The average nucleotide compositions and cumulative lengths of the 18S rRNA stems and loops, respectively, were also calculated.

3.3.2 Growth temperature

Optimal or near optimal growth temperatures of 9168 eubacterial and archaebacterial species entries were kindly provided by Dr. Manfred Kracht of German Collection of Microorganisms and Cell Cultures (http://www.dsmz.de). Many species are listed more than once in the database, sometimes with different growth temperature. For example, there are 101 temperature entries for *Bacillus sp.*, 82 of which have a temperature of 30°C (Fig. 3.2). In other cases the temperatures are all the same or similar. For simplicity the temperature of first entry of the same species was assigned for that species, which leaves temperature data of 3630 unique species. We also calculated average temperature for the same species if there were multiple entries. In the case of *Bacillus sp.* for instance,

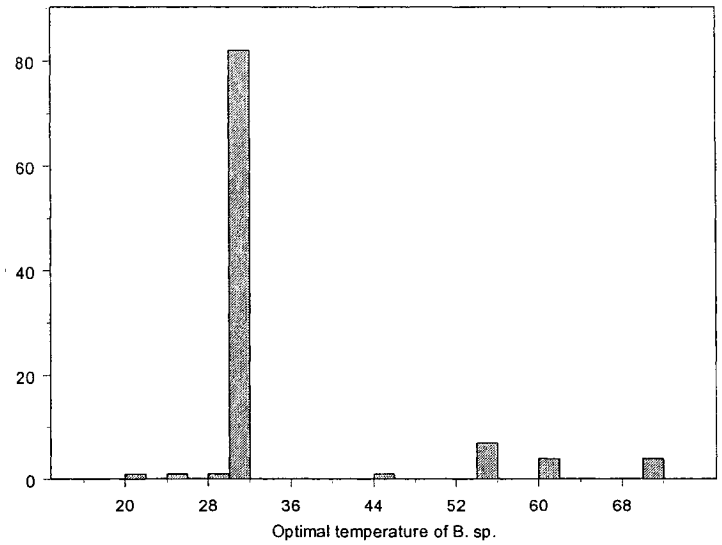


Fig. 3.2 Distribution of optimal growth temperature of *Bacillus sp.* The X axis is the temperature and Y axis is the number of entries.

the average temperature is 34.49 °C. The correlation of the two temperature data sets is 0.98. There is no marked difference in subsequent analyses when using the two temperature data sets. Here we used the first temperature data, i.e., temperature of first entry is assigned to a species when there are multiple entries for it. The temperature data were then merged with the rRNA data and 1673 species that have both temperature and rRNA data (GC content and gene length) were obtained. The temperature distribution (Fig. 3.3) shows that the majority of the species have a temperature of 28, 30 or 37 °C. Of the 1673 species, 1573 are bacterial and the other 100 are archaeal; 1530 are mesophilic (temperature $< 45°C$) and the other 143 are thermophilic (temperature $\geq 45°C$). In addition to the two-way separation of mesophiles and thermophiles we also separated the species into three temperature groups: less than 40°C (1522 mesophilic species), between 40°C and 75 °C (114 moderately thermophilic species), and higher than 75 °C (37 hyperthermophilic species).

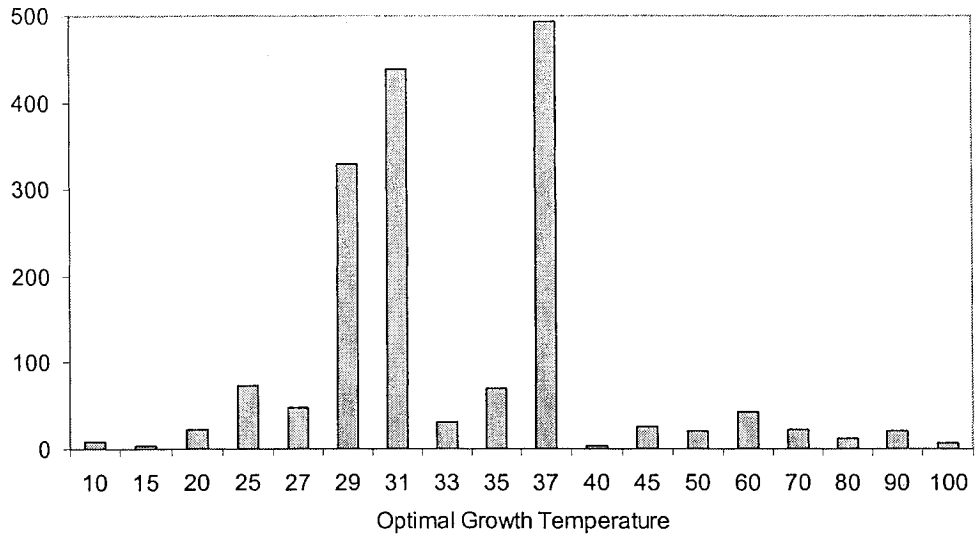


Fig. 3.3 Distribution of optimal growth temperature of 1673 prokaryotic species. The X axis is the temperature in °C and the Y axis is the number of the species.

3.3.3 Statistical analyses

We used the nonparametric Wilcoxon rank sum tests to compare average rRNA GC content and length between two temperature groups such as mesophiles and thermophiles in bacteria. The Kruskal-Wallis rank sum tests were used to compare the data among three temperature groups, such as mesophilic, moderately thermophilic and hyperthermophilic bacteria. These tests were performed using the statistics package SYSTAT version 10 (SPSS Science, 2000). To reduce phylogenetic dependence, we extracted those genera that have at least one mesophilic species and one thermophilic species in the assembled rRNA-temperature data set. Paired *t*-tests were used to compare the means of GCrRNA and rRNA length of the mesophilic species and thermophilic species for each of these genera. For the archaea we obtained phylogenetic independent contrasts for temperature, rRNA GC content and length. The contrasts were calculated using the CONTRAST program (Felsenstein, 1985) implemented in PHYLIP 3.6 (beta release) (Felsenstein, 2004b) based on two recently published archaeal trees (Fig. 3.4; Brochier et al., 2004). One of them (Fig. 3.4a) was constructed from the concatenated sequences of transcription proteins (such as RNA polymerases) of 20 archaeal species whose complete genomes have been sequenced. Another tree (Fig. 3.4b), the translation proteins tree, was derived from a concatenation of ribosomal proteins of 18 archaeal

species whose whole genome sequences are known. The protein trees rather than an rRNA tree were used here so that the rRNA contrasts were not derived from the rRNA tree itself.

16S rRNA sequences for the 20 archaeal species were extracted from the ssu rRNA database (Wuyts et al., 2002) or GenBank if they were not available in the rRNA database. For *Methanosarcina acetivorans*, only a partial sequence of 16S rRNA (MESRR16SA, 1426 bp) was retrieved. For *Ferroplasma acidarmanus*, a partial sequence (AF145441) of 849 bp was found in GenBank and since it is too short, the sequence (AY222042) of its close relative *F. acidiphilum* was used, although it was also annotated as partial (1394 bp) in GenBank.

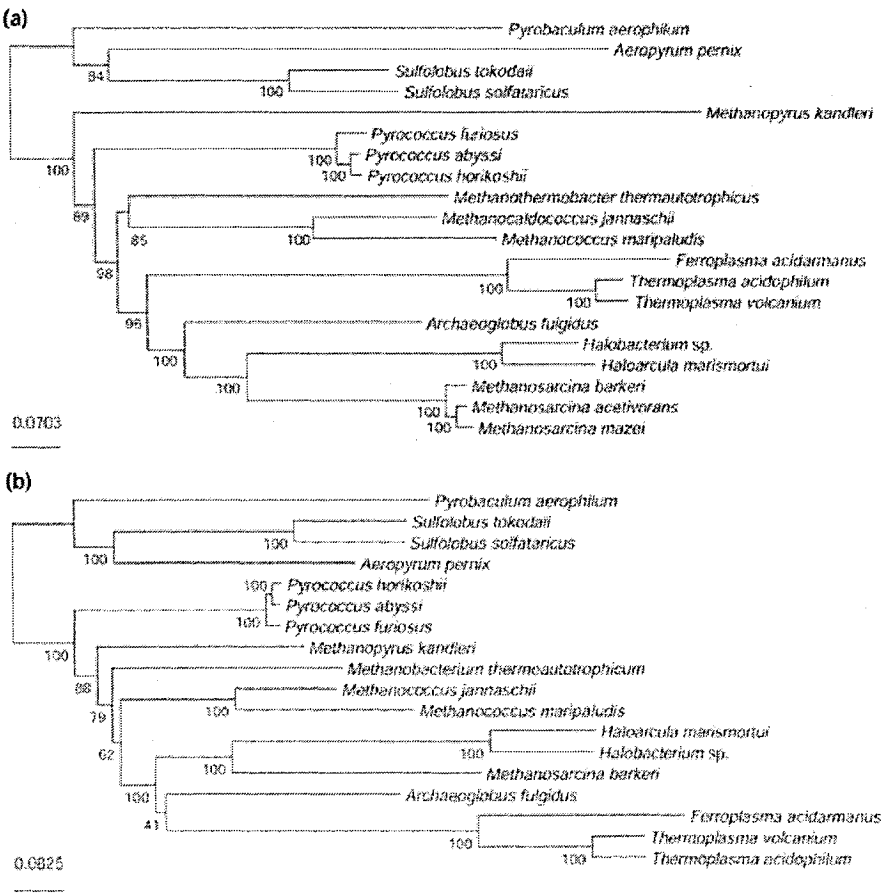


Fig. 3.4 Archaeal phylogenetic trees based on transcription proteins (a) and translation proteins (b) using unrooted maximum likelihood method (Brochier et al. 2004). These two trees are used for deriving phylogenetic independent contrasts for the traits of the archaeal species.

3.4 Results

3.4.1 Nucleotide composition and sequence length of 16S rRNA in prokaryotes

Table 3.1A shows average GC content and length of 16S rRNA, its stems and loops in the low, medium and high temperature groups for the 1573 bacterial species (1461 species in the group of less than 40 °C, 106 species in 40-75 °C group and 6 species in higher than 75 °C group). The rRNA GC and especially the stem GC are highest in the hyperthermophiles, lowest in the mesophiles and moderate thermophiles in the middle.

Table 3.1 Average GC content and sequence length of 16S rRNA stems and loops for mesophilic (< 40°C), moderately thermophilic (40-75 °C) and hyperthermophilic (≥ 75°C) bacteria (A) and archaea (B).

A) Bacteria (n = 1573)	Sequence composition (GC%)			Sequence length (bases)		
	rRNA	Stems	Loops	rRNA	Stems	Loops
Mesophile (n=1461)	55.24±2.78**	66.1 ± 0.11	40.1 ± 0.04	1522.5±24.10	885.2±0.52	637.3±0.39
Thermophile (n=106)	58.19±2.73	70.1± 0.37	41.6 ± 0.17	1535.5±46.25	894.9±2.88	640.6±2.44
Hyperthermophile (n=6)	62.25 ± 2.58	76.5 ± 1.55	41.2 ± 0.35	1556.5±14.08	922.2±6.87	634.3±4.14
p-value*	<0.000001	<0.000001	<0.000001	0.0003	<0.000001	0.74

B) Archaea (n = 100)	Sequence composition (GC%)			Sequence length (bases)		
	rRNA	Stems	Loops	rRNA	Stems	Loops
Mesophile (n=61)	56.14±1.64	67.4 ± 0.26	38.97 ± 0.21	1471.5±10.19	881.3±1.83	590.2± 1.17
Thermophile (n=8)	58.49±3.29	70.8 ± 1.62	39.2 ± 0.57	1482±12.5	891.1± 4.55	590.9± 2.38
Hyperthermophile (n=31)	65.8±1.72	81.0 ± 0.41	42.1 ± 0.25	1492.6±12.28	903.2± 3.03	589.4± 2.01
p-value*	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001	0.58

*p-value for Kruskal-Wallis rank sum test for variables in each column.
** mean ± standard errors

The rRNA length and the cumulative length of the stems are also in the order of hyperthermophiles > moderate thermophiles > mesophiles. Both GC content and the cumulative length of the loops are largest in the moderate thermophiles, but the differences are relatively small among the three temperature groups. The Kruskal-Wallis rank sum tests indicated the differences in mean GC content among the three groups are very significant for rRNA GC, stem GC and loop GC (p-value < 0.000001). The tests for the differences in mean rRNA length and cumulative stem length are also significant ($p = 0.0003$ and $p < 0.000001$, respectively), but the test for difference in loop length is insignificant.

Table 3.1B shows analog results for 100 aracheal species (61 species in the group of less than 40 °C, 8 species in 40-75 °C group and 31 species in higher than 75 °C group). rRNA GC, stem GC, rRNA length and cumulative stem length are also in the order of hyperthermophiles > moderate thermophiles > mesophiles. The Kruskal-Wallis rank tests are all very significant. The loop GC is also in this order and the Kruskal-Wallis rank test is also significant but the difference is much smaller than seen in the stems or the whole rRNA. This is partially due to the ambiguity in assigning rRNA nucleotides to stem and loop structures. The cumulative loop length is relatively longer in the moderate thermophiles and shorter in the hyperthermophiles but the difference between the groups is very small such that the Kruskal-Wallis rank test is insignificant, indicating there is no difference in the loop length among the three temperature groups of archaeal species.

These results on bacteria and archaea confirmed our prediction that (hyper)thermophiles require not only high GC but also longer rRNA, and the increases in both GC and length are concentrated in the stem regions. Both loop GC and length are however very stable. Table 3.1A and B also indicate that bacterial rRNAs are much longer than the archaeal ones for the corresponding temperature groups and the difference is mainly in the loop length, this is because the former have two extra hairpin structures in the molecule (Wuyts et al., 2002).

3.4.2 Genus level comparisons

The 1573 bacterial species in the 16S rRNA data set contain 444 genera, of which 19 genera have at least one mesophilic species and one thermophilic species. The means of the optimal growth temperature, the rRNA GC content and length of the mesophilic species and thermophilc species for each of the 19 genera were calculated (Table 3.2). The mean rRNA GC contents are higher in the thermophilic species than in the mesophilic species ($p = 0.0003$ for a paired two-tail *t*-test). The mean rRNA lengths are

Table 3.2 Average 16S rRNA GC content (%) and length (bases) of mesophilic species (Meso) and thermophilic species (Thermo) in the same genus.

Genus*	Temperature (°C)**		Sequence Composition (GC%)		Sequence Length (bases)	
	Meso	Thermo	Meso	Thermo	Meso	Thermo
Actinomadura (18, 1)	30.74	55	59.77	60.22	1507	1516
Actinopolyspora (1, 1)	37	45	60.88	61.43	1546	1537
Amycolatopsis (7, 1)	28.29	45	58.7	59.1	1510	1517
Bacillus (46, 7)	30.03	55	54.71	58.17	1541	1540
Brevibacillus (6, 2)	31.65	47.5	55.4	56.26	1523	1541
Clostridium (101, 4)	34.57	57.34	52.96	53.63	1514	1553
Deinococcus (1, 2)	30	47.5	55.53	58.6	1502	1498
Desulfotomaculum (6, 8)	34.08	56.25	53.97	57.31	1530	1563
Lactobacillus (45, 1)	32.97	45	52.39	53.73	1561	1560
Mycobacterium (43, 2)	35.6	45.5	58.07	59.4	1516	1522
Porphyrobacter (1, 2)	30	45	53.65	53.92	1480	1477
Pseudonocardia (7, 1)	28.64	45	59.45	59.18	1515	1522
Rubrobacter (1, 1)	37	50	57.31	61.71	1539	1544
Saccharomonospora(4, 2)	29.42	47.43	59.32	60.31	1516	1512
Saccharopolyspora (5, 1)	28	51.88	58.88	59.29	1515	1517
Spirochaeta (7, 1)	32.71	65	54	60.66	1563	1532
Streptomyces (67, 5)	28.26	46.5	58.95	60.02	1520	1521
Sulfobacillus (1, 1)	35	45	59.22	63.6	1525	1538
Thermoactinomyces (1, 5)	35	51.08	56.14	58.76	1536	1517
Mean ± S.D.	32.05±0.7	49.8±1.3	56.8 ± 0.61	58.7 ± 0.62	1524 ± 4.6	1528 ± 4.8

*The first number in the bracket is the number of mesophilic species and the second number is the number of thermophilic species in the genus.

** Average growth temperature of the mesophilic species and thermophilic species in the genus.

also longer in the thermophilic species than in the mesophilic species, but the difference is not significant (p = 0.34 for a paired *t*-test). However, if we separate the rRNAs into stems and loops according to their secondary structure, the mean cumulative stem lengths are longer in the thermophilic species than in the mesophilic species (p = 0.018, paired *t*-test), while the difference in loop length is not significant (p = 0.62, paired *t*-test) for the 19 genera. The GC contents of the stems and loops are higher in the thermophilic species than in the mesophilic species and the difference are significant in both cases (p < 0.00001) (Table 3.3).

Table 3.3 Average GC content (%) and cumulative length (bases) of stems and loops of 16S rRNA of mesophilic species (Meso) and thermophilic species (Thermo) in the same genus.

	Sequence Composition (GC%)				Sequence Length (bases)			
	Stems		Loops		Stems		Loops	
Genus	Meso	Thermo	Meso	Thermo	Meso	Thermo	Meso	Thermo
Actinomadura	73.09	73.56	41.47	41.91	873	876	633	640
Actinopolyspora	73.39	74.94	42.79	42.58	909	916	637	621
Amycolatopsis	71.21	71.77	41.56	41.53	870	883	640	634
Bacillus	64.6	68.9	40.58	42.77	907	903	634	636
Brevibacillus	65.49	66.62	40.79	41.16	903	912	620	628
Clostridium	63.07	63.19	39.1	40.5	872	899	642	654
Deinococcus	66.59	70.21	40.13	42.06	874	878	628	620
Desulfotomaculum	64.08	68.93	39.76	41.77	895	901	635	661
Lactobacillus	61.49	62.9	39.37	40.75	913	911	648	649
Mycobacterium	70.25	72.07	41.21	41.92	884	885	632	637
Porphyrobacter	65.21	65.43	38.23	38.24	845	852	635	625
Pseudonocardia	72.22	72.34	41.94	40.97	874	882	641	640
Rubrobacter	69.82	76.15	40.25	42.1	888	888	651	656
Saccharomonospora	71.85	73.06	42.24	42.84	865	867	651	645
Saccharopolyspora	70.89	71.87	42.38	41.82	873	865	641	652
Spirochaeta	65.8	75	38.26	40.22	891	920	671	612
Streptomyces	71.36	72.87	41.73	42.01	885	891	634	629
Sulfobacillus	69.5	74.29	44.87	48.5	887	884	638	654
Thermoactinomyces	67.39	70.82	40.26	42.02	897	898	639	618
Mean ± S.E.	68.3±0.84	70.8±0.89	40.9±0.38	41.9±0.44	884.5±4.0	890.1±4.20	639.5±2.45	637.4±3.33

The 100 archaeal species contain 45 genera. Only one of them (*Methanosarcina*) has both mesophilic and thermophilic species. In that genus, the mean rRNA length is

longer and the mean rRNA GC is lower in the thermophilic species than in the mesophilic species. As there is only one such genus in the archaeal data the statistical significance cannot be inferred. To see further the relationship of the rRNA length and GC content with temperature we conducted independent contrasts for the archaeal species.

3.4.3 Phylogenetic-based comparison

The Contrast program in the PHYLIP package was used to derive phylogenetic independent contrasts for rRNA GC, length and temperature simultaneously, based on the two archaeal phylogenetic trees shown in Fig. 3.4a (transcription protein tree) and Fig. 3.4b (translation protein tree), respectively. Table 3.4 lists the correlation coefficients of rRNA GC content and length with optimal temperature for the original data and for the contrasts based on the transcription tree and translation tree, respectively. This indicates the correlation between rRNA GC and temperature remains very strong for both translation tree and transcription tree based contrasts, as does in the original data. The correlation between rRNA length contrast and temperature contrast is also very strong and even higher than that for the original data. It is weaker, however, in the contrasts based on the transcription tree.

Table 3.4 Correlation coefficients of rRNA GC content and length with optimal growth temperature for original data (20 species), contrasts based on the translation tree (18 species) and contrasts based on the transcription tree (20 species).

	GC and temperature	Length and temperature
Original data	0.90 (p = 5.59e-8)	0.79 (p = 3.77e-5)
Translation tree based contrasts	0.85 (p = 6.74e-5)	0.84 (p = 0.0001)
Transcription tree based contrasts	0.84 (p = 5.77e-6)	0.44 (p = 0.043)

Fig. 3.5 shows the regression of the rRNA GC contrasts on temperature contrasts based on the transcription tree (slope = 0.14, $p < 0.00001$ for the null hypothesis that regression slope = 0). The regression R^2 is 0.71, indicating a large proportion of variance in the GC content can be explained by the temperature. The regression of the rRNA length contrast on temperature contrast (Fig. 3.6) has a regression R^2 of 0.19 ($p = 0.043$),

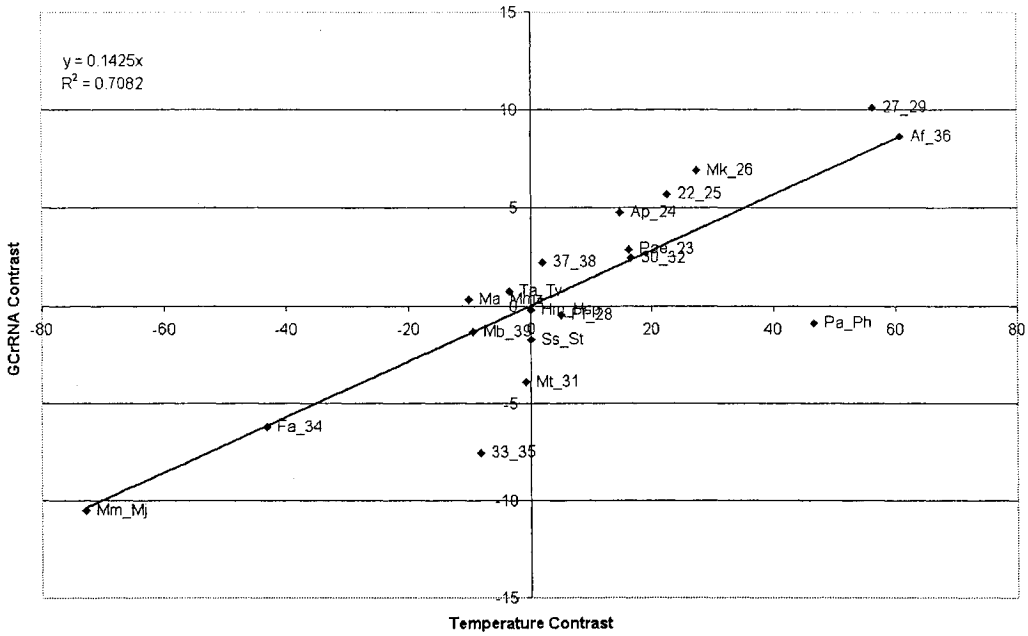


Fig. 3.5 Regression of rRNA GC content contrasts on temperature contrasts based on the transcription tree of 20 archaeal species (Fig. 3.4a; Brochier et al., 2004). Each contrast is a difference in traits (i.e., rRNA GC and temperature) between two species or between one species and one node, or between two nodes. The nodes are in numbers and the species are in two letters, which has the following correspondence to the species name. Pae *Pyrobaculum aerophilum*; Ap *Aeropyrum pernix*; Ss *Sulfolobus solfataricus*; Sk *Sulfolobus tokodaii*; Mk *Methanopyrus kandleri*; Pf *Pyrococcus furiosus*; Pa *Pyrococcus abyssi*; Ph *Pyrococcus horikoshii*; Mt *Methanobacterium thermoautotrophicum*; Mm *Methanococcus maripaludis*; Mj *Methanococcus jannaschii*; Fa *Ferroplasma acidiphilum*; Ta *Thermoplasma acidophilum*; Tv *Thermoplasma volcanium*; Af *Archaeoglobus fulgidus*; Hm *Haloarcula marismortui*; Hsp *Halobacterium sp.*; Mb *Methanosarcina barkeri*; Ma *Methanosarcina acetivorans*; Mmz *Methanosarcina mazei*.

indicating most of the variance of rRNA length cannot be explained by the temperature. This was caused by the two outlier points associated with three *Methanosarcina* species (*M. acetivorans*, *M. mazei* and *M. barkeri*) (Ma_Mmz and Mb_node39 in Fig. 3.6). *M. acetivorans* rRNA has only 1426 bases which is partial while *M. mazei* rRNA has a full length of 1474 bases, thus causing a huge contrast. It further causes a large contrast for the node (node 39) consisting of *M. acetivorans* and *M. mazei* and the species *M. barkeri*. Indeed, when the two outlier points were removed the correlation between the rRNA length contrasts and temperature contrasts becomes stronger (R = 0.88, p = 1.76e-5) than that seen in the original data (R = 0.79). This is also demonstrated in Fig. 3.7, the regression of the rRNA length contrast on temperature contrast based on the transcription

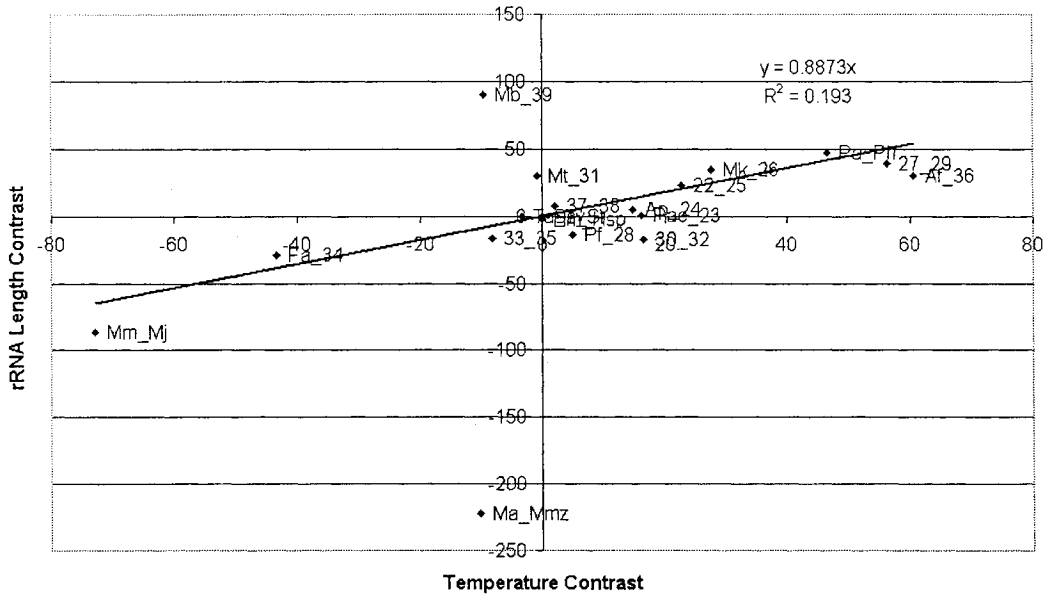


Fig. 3.6 Regression of rRNA length contrasts on temperature contrasts based on the transcription tree of 20 archaeal species (Fig. 3.4a; Brochier et al., 2004). Two outliers points are Ma-Mmz and Mb-node 39. The nodes are in numbers and the species are in two letters. See legends in Fig. 3.5 for the species abbreviations.

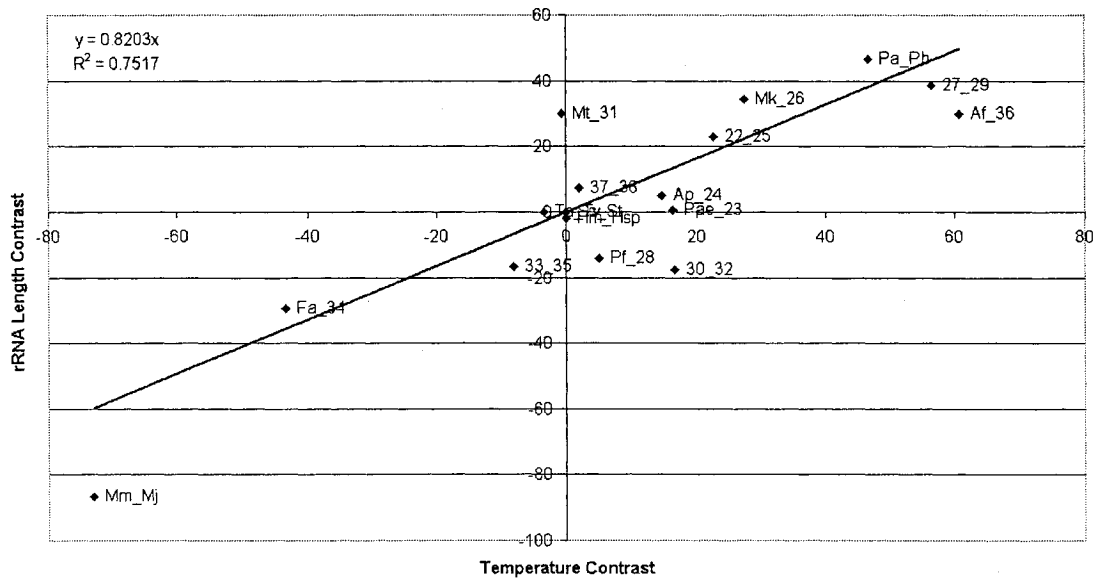


Fig. 3.7 Regression of rRNA length contrasts on temperature contrasts based on the transcription tree (Fig. 3.4a; Brochier et al., 2004), when the two *Methanosarcina* associated outlier points in Fig. 3.6 (Ma_Mmz, contrast between *M. acetivorans* and *M. mazei*, and Mb_39, contrast between the node of the two *Methanosarcina* and the species *M. barkeri*) were removed. The nodes are in numbers and the species are in two letters. See legends in Fig. 3.5 for the species abbreviations.

tree when the two *Methanosarcina* outlier points were removed ($p < 0.0001$ for the null hypothesis that regression slope = 0).

The translation tree (Fig. 3.4b) does not contain *M. acetivorans*, *M. mazei*, thus posing no such problem arising from the partial sequence of *M. acetivorans*. Figures 3.8 and 3.9 are the regressions of GC rRNA contrast and rRNA length contrast, respectively, on temperature contrast based on the translation tree. Both indicate that most of the variances in rRNA GC and in rRNA length can be explained by the temperature.

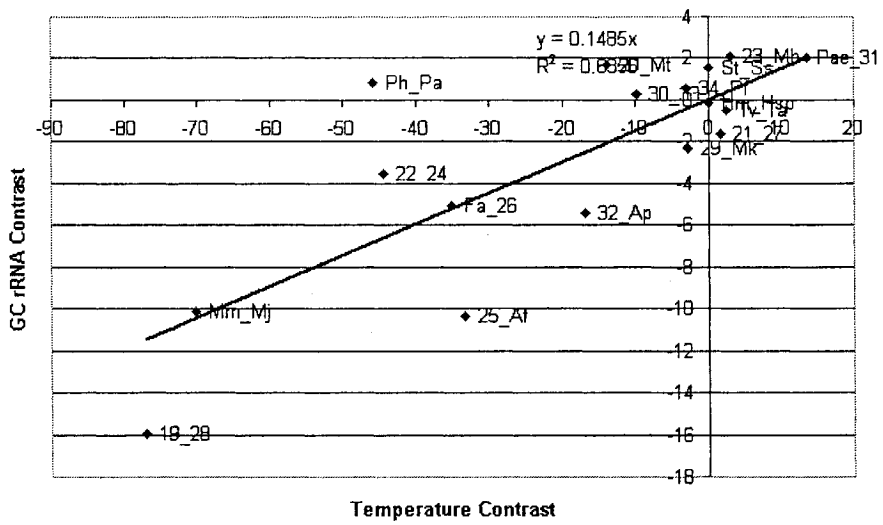


Fig. 3.8 Regression of GC rRNA contrasts on temperature contrasts based on the translation tree of 18 archaeal species (Fig. 3.4b). The nodes are in numbers and the species are in two letters.

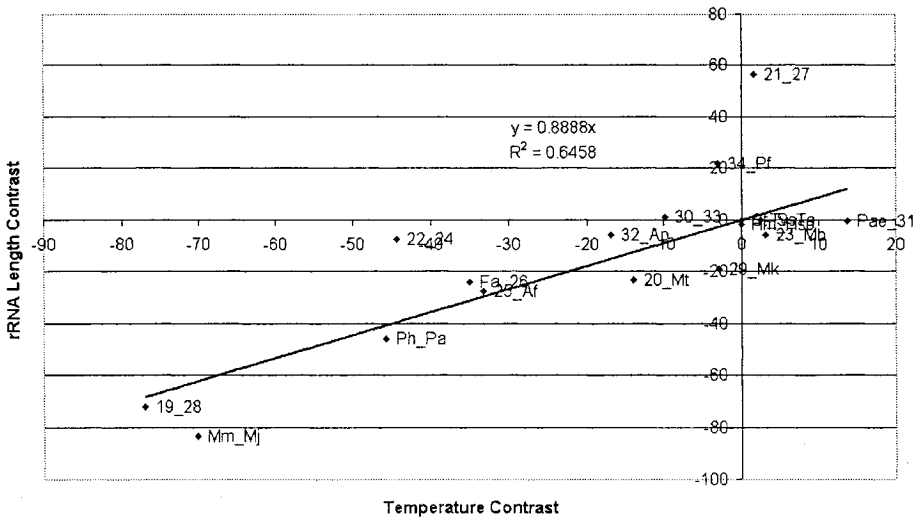


Fig. 3.9 Regression of rRNA length contrasts on temperature contrasts based on the translation tree (Fig. 3.4b). The nodes are in numbers and the species are in two letters.

3.4.4 Nucleotide composition and length of vertebrates 18S rRNA

Having demonstrated the relationship between bacterial 16S rRNA and optimal growth temperature, it is interesting to see the rRNA-temperature relationship in vertebrates, as the latter can be broadly separated warm-blooded vertebrates, including mammals and birds, and cold-blooded vertebrates, including fish, amphibians and reptiles. The difference in average 18S rRNA GC between the mammals (4 species) and birds (34 species) is insignificant ($p = 0.60$, Wilcoxon rank-sum test). It is also insignificant among the cold-blooded vertebrates (23 fish, 18 amphibian and 5 reptile species) ($p = 0.27$, Kruskal-Wallis rank-sum test). However both mammal and birds have higher average GC than the cold-blooded vertebrates (Fig. 3.10) and the difference in GC content between the warm and cold-blooded vertebrates is significant ($p = 0$, Wilcoxon rank-sum test).

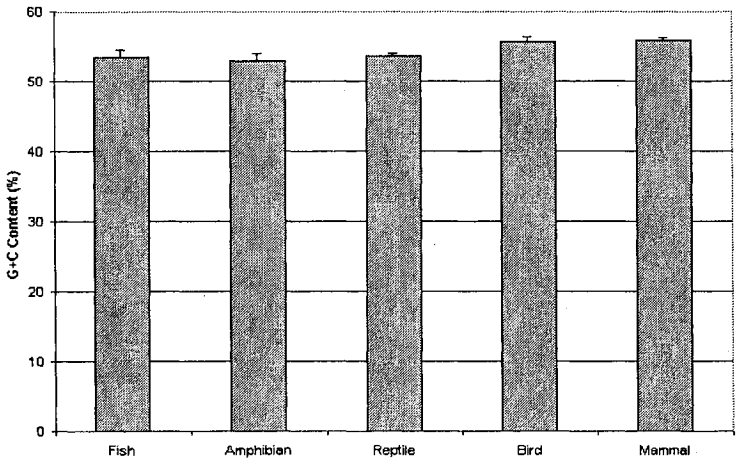


Fig. 3.10 Average 18S rRNA G+C content and standard deviations (error bars) of 84 vertebrate in five groups. Birds and mammals have higher amount of G+C than the other vertebrates.

The average length 18S rRNA is higher in mammals than in birds (Fig. 3.11) and the difference is highly significant ($p = 0$, Wilcoxon rank-sum test). The average length of the rRNA is not different among fish, amphibian and reptiles ($p = 0.10$, Kruskal-Wallis rank-sum test). The average pooled length between the warm and cold-blooded vertebrates is marginally different ($p = 0.049$, Wilcoxon rank-sum test). Figure 3.11 also shows that the average 18S rRNA length of the mammals is longest and that of the reptiles is shortest among the five vertebrate groups.

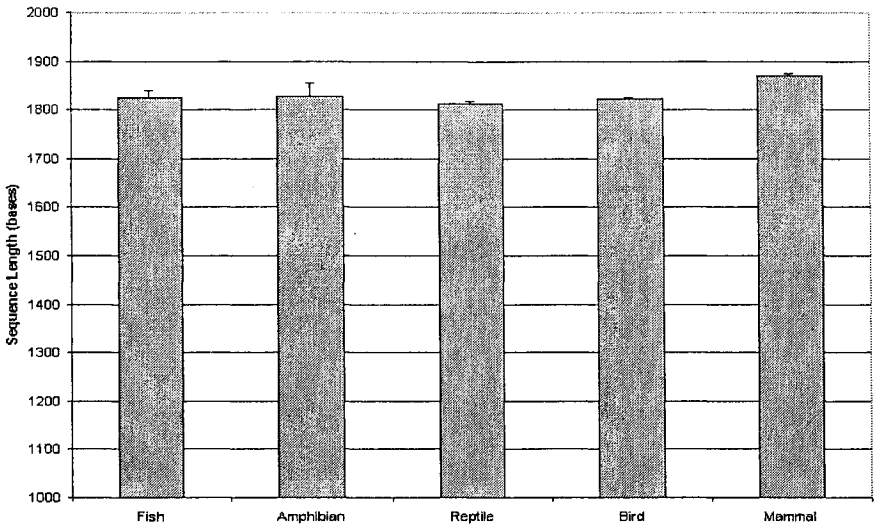


Fig. 3.11 Average 18S rRNA length and standard deviation (error bars) of 84 vertebrates in five groups. Mammals have the longest rRNA while reptiles and birds have the shortest rRNA.

Figures 12 and 13 show individual nucleotide compositions of the vertebrate 18S rRNA stems and loops, respectively. The rRNA stems of both mammals and birds have higher amount of G than that of cold-blooded vertebrates. In all five groups of vertebrates the amounts of G and C are not equal, nor are the amounts of T and A, indicating the base

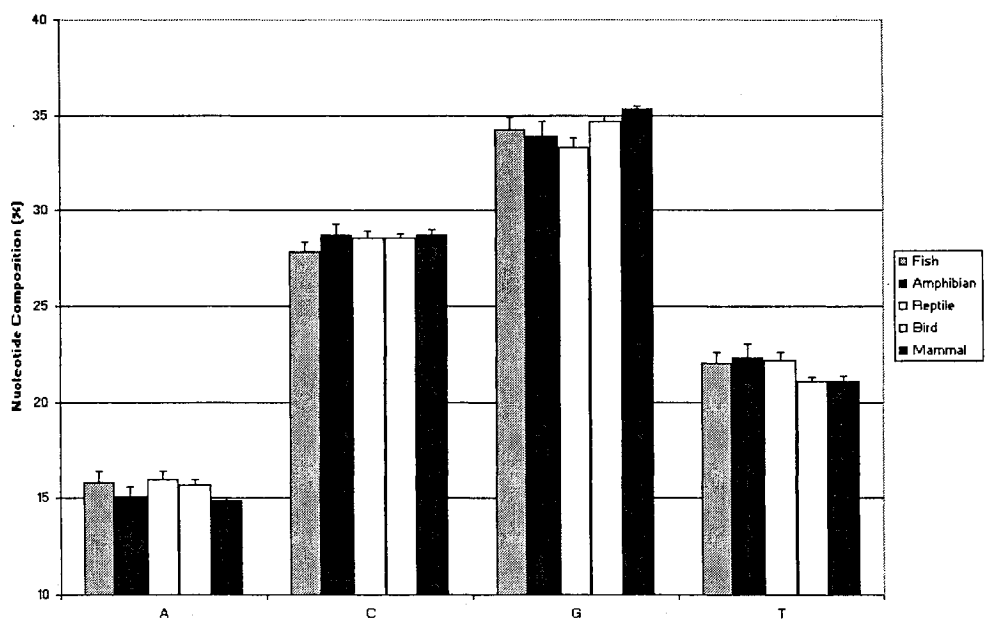


Fig. 3.12 Average nucleotide composition and standard deviations (error bars) of 18S rRNA stems of 84 vertebrate in five groups.

parity rule for RNA helical structures is not held for the rRNA stems. This is because the stems frequently contain non-canonical bases pairs such as G:T pairs in addition to the A:T and G:C pairs. The rRNA loops of all five groups of vertebrates have very high amount of A (the mean amount greater than 30% in each group). Accordingly, the 18S rRNA loops have very low amounts of C and G, especially in the cold-blooded vertebrate groups.

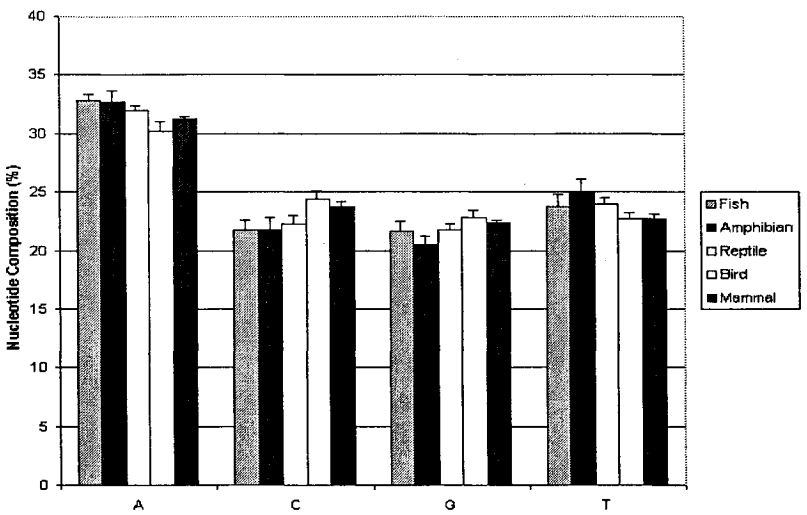


Fig. 3.13 Average nucleotide composition and standard deviations (error bars) of 18S rRNA loops of 84 vertebrate in five groups. Adenine is in highest amount for all species.

3.5 Discussion

Prokaryotes: Using the methods of genus level comparison, and also phylogenetic independent contrasts, we confirmed the previous findings of a strong positive correlation between the GC content of rRNA and environmental growth temperature, among both eubacterial and archaeal species. Therefore, we can conclude that the relationship between the nucleotide content of structural RNAs and environmental growth temperature is not due to phylogenetic history, but reflects a repeated selective response to elevated environmental temperature. We also found a positive relationship between rRNA length and growth temperature and this length difference occurred primarily in the paired regions of the molecule. Taken together, these results indicate that prokaryotic rRNAs respond to increased environmental growth temperature by increasing the

structural stability of their rRNAs. This is achieved both by increasing the GC content and the length of the paired regions. Both of these factors (increased GC and increased length of paired regions) increase the number of hydrogen bonds between the paired strands. Thus it is reasonable to interpret these changes as adaptations to growth at high temperature.

Vertebrates: Among vertebrates, a similar correlation was observed, in that 18S rRNAs of the warm-blooded animals (birds and mammals) have higher G+C contents than those of the cold-blooded animals (fish, amphibians and reptiles). Also, the 18S rRNA length of mammals is the longest among the five vertebrate groups. When examining the data in more detail, however, we notice that the differences are not large and that they are not concentrated in the paired regions of the molecule (Fig. 3.14). At first glance, this seems

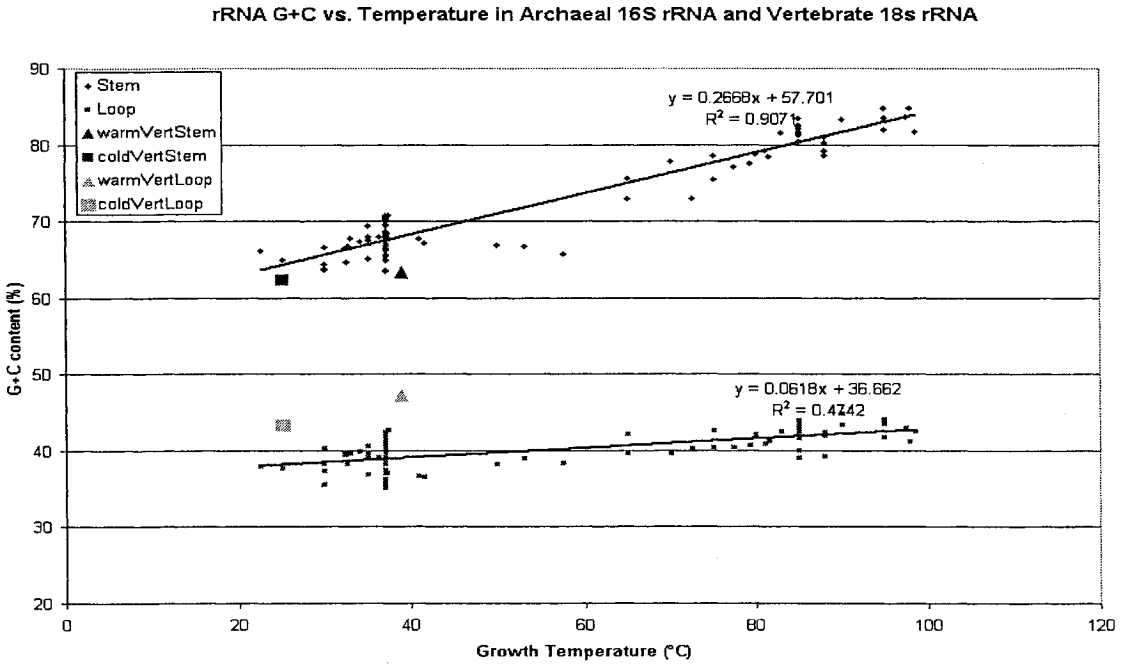


Fig. 3.14 (upper panel) The GC content of 16s rRNA stems (shown in small diamonds) increases rapidly with optimal growth temperature in 100 archaeal species, while the average stem GC of 38 warm-blooded vertebrates (shown in large triangle) have little increase over the average stem GC of 46 cold-blooded vertebrates (in large square). (lower panel) The GC content of the archaeal loops (in small squares) shows little increase with the temperature while the average loop GC of warm-blooded vertebrates (in large triangle) is higher than that of cold-blooded vertebrates (in large square). The preferred body temperature of the cold and warm-blooded vertebrates is set to be 25 and 39 °C, respectively.

to contradict the findings on prokaryotes, but when we consider that 37-39 °C (body temperature of mammals are about 37 °C and that of birds are about 39 °C) is not even within the "moderately thermophilic" range, the result is not at all surprising.

Despite the lack of temperature-induced differences between the rRNA sequences of warm-blooded and cold-blooded vertebrates, these sequences do illustrate many of the same features as the prokaryotic 16S rRNAs. For instance, the paired regions are relatively GC-rich while the rRNA loops of all five groups of vertebrates have a very high amount of adenine. This is reminiscent of high amount of A in prokaryotic 16S rRNA loops in both mesophilic and thermophilic prokaryotes (Wang & Hickey, 2002) and also in the current data of 1673 bacterial and archaeal species (data not shown). In fact, this compositional bias also exists in other eukaryotic species, including protists, fungi, invertebrates and plants (data not shown). This finding of a strong bias of adenine in 16S and 18S rRNA loops is consistent with a previous comparative analysis of the covariation-based structure model of 16S and 23S rRNAs, which revealed that the majority of adenines are unpaired, while the majority of the G, C and U bases are paired (Gutell et al., 2000).

The lowest amount of A in both 16S rRNA and 18S rRNA stems can be explained by the fact that all stems have higher G+C than A+T and G:U pairs are also frequent in stem regions, which causes fewer A than T (U) in rRNA stem regions. The elevated frequency of A in loop regions might be explained by the availability of adenines in the cellular environments; as the stems use fewer adenines there is a higher concentration of As left in the environment for forming loops. But the rRNA stems may not affect the overall nucleotide pool significantly. Alternatively, the unpaired adenines may contribute to the structural stability of the molecule. Based on the following three facts: (1) unpaired adenines in asymmetrical loops are more destabilizing than those in symmetrical loops, (2) the majority of adenines in the unpaired regions occur in asymmetric loops, and (3) the majority of these unpaired adenosine nucleotides are base-paired, albeit in an irregular manner, Gutell et al. (2000) speculated that:

"These destabilizing, asymmetrically placed adenines are a significant component in the transition from secondary to tertiary RNA structure. The destabilizing effects of these adenines in secondary structure, coupled with the need for an RNA molecule to

adopt its minimal energetic state, suggest that these abundant adenosine nucleotides will actively seek out energetically stabilizing tertiary interactions and, in the process, form a three-dimensional RNA molecule."

Biases in favor of increased adenine usage have also been observed in mitochondrial coding sequences (Xia, 1996), in the genomes of obligatory pathogenic bacteria, plasmids and bacterial phages (Rocha & Danchin, 2002) and endosymbiont bacteria and fungi (Woolfit & Bromham, 2003), and in the coding sequences of thermophilic bacteria (Singer & Hickey, 2003). The structural or functional roles of these biased adenosine nucleotides in these sequences remain to be revealed.

Chapter 4

Mutational bias affects protein evolution in flowering plants*

4.1 Abstract

Amino acid sequences from several thousand homologous gene pairs were compared for two plant genomes, *Oryza sativa* and *Arabidopsis thaliana*. The *Arabidopsis* genes all have similar GC contents, whereas their homologs in rice span a wide range of GC levels. The results show that those rice genes that display increased divergence in their nucleotide composition (specifically, increased GC content) showed a corresponding, predictable change in the amino acid compositions of the encoded proteins relative to their *Arabidopsis* homologs. This trend was not seen in a "control" set of rice genes that had nucleotide contents closer to their *Arabidopsis* homologs. In addition to showing an overall difference in the amino acid composition of the homologous proteins, we were also able to investigate the biased patterns of amino acid substitution since the divergence of these two species. We found that the amino acid exchange matrix was highly asymmetric when comparing the high GC rice genes to their *Arabidopsis* homologs. Finally, we investigated the possible causes of this biased pattern of sequence evolution. Our results indicate that the biased pattern of protein evolution is the consequence, rather than the cause, of the corresponding changes in nucleotide content. In fact, there is an even more marked asymmetry in the patterns of substitution at synonymous nucleotide sites. Surprisingly, there is a very strong negative correlation between the level of nucleotide bias and the length of the coding sequences within the rice genome. This difference in gene length may provide important clues about the underlying mechanisms.

* Main part of the contents has been published in Wang H-C., G.A.C. Singer and D.A. Hickey (2004) *Molecular Biology and Evolution* 21: 90-96.

4.2 Introduction

Differences in GC content among genomes have been intensively studied and wide variations have been noted, both among entire genomes, and among genes within genomes (Li, 1997; Karlin, Campbell & Mrazek, 1998; Gautier, 2000). The differences in nucleotide content between genomes have been shown to cause concomitant changes in the amino acid compositions of the encoded proteins (Collins & Jukes, 1993; Foster, Jermiin & Hickey, 1997; Lobry, 1997; Wilquet & Van de Casteele, 1999; Singer & Hickey, 2000; Kreil & Ouzounis, 2001). Most of these previous studies were based primarily on prokaryotic genomes, due to the lack of large-scale genomic data for plants and animals. Such data are now becoming available, however. The recent availability of genomic data for multicellular plants and animals not only allows us to extend previous studies to the genomes of multicellular eukaryotes, but it also enables us to trace the patterns of nucleotide and amino acid substitution between lineages that have well-defined evolutionary relationships. Therefore, we not only see the end results of evolutionary changes between genomes, but we can also trace the paths by which these changes took place.

In this study, we compared homologous gene pairs from two species of flowering plants, rice and *Arabidopsis*. Because these two species diverged less than two hundred million years ago, many homologous sequences from the two genomes are unambiguously alignable. Moreover, the level of amino acid sequence divergence between homologous proteins is relatively low, allowing us to gauge the patterns of amino acid substitution. Finally, there is a wide variation in the nucleotide contents of the rice genes: some rice genes closely resemble their *Arabidopsis* homologs in GC content, while others have significantly elevated levels of GC relative to their homologs (Carels & Bernardi, 2000). Since all of the genes diverged from their common ancestral sequences at the same point in evolutionary time, this provides us with a "controlled" evolutionary experiment, enabling us to do a comparative study of two sets of rice genes that are evolving under contrasting evolutionary constraints.

4.3 Materials and Methods

4.3.1 Sources of sequence data

Protein sequences from *O. sativa* were obtained from the Gramene database (Ware et al. 2002) (ftp://www.gramene.org/pub/gramene/protein/sequence/rice_sptrembl.fa). This database contained 8985 sequences as of May 2002. From the protein sequence identifiers we first obtained corresponding EMBL accession numbers by searching SwissProt (Bairoch & Apweiler 2000), then extracted corresponding EMBL sequence records (Stoesser et al. 2002). From the EMBL records we wrote a program to extract coding sequences and 9916 coding sequences were obtained. A total of 443 sequences were shorter than 75 codons and were excluded from the analysis. The remaining sequences were subjected to a codon integrity check using CodonW (http://www.molbiol.ox.ac.uk/cu/), and we further cleaned the data by removing redundant sequences. The final data set of *O. sativa* coding sequences contains 7886 non-redundant sequences. Using EMBOSS/transeq (Rice, Longden & Bleasby 2000) to translate the file, we generated a corresponding non-redundant protein sequence file.

A total of 26,178 protein coding sequences from *A. thaliana* (from five chromosomes) were downloaded from National Center for Biotechnology Information (NCBI) FTP server (ftp://ftp.ncbi.nih.gov/genbank/genomes/A_thaliana/). After passing the sequences to CodonW for codon integrity check, and removing genes shorter than 75 codons a total of 25,625 *Arabidopsis* coding sequences remained for analysis. Protein sequences of *Arabidopsis* were also obtained by translating the coding sequences using EMBOSS/transeq program.

4.3.2 Identification and comparison of homologous sequences

Homologous protein pairs between *O. sativa* and *A. thaliana* were identified by performing BLASTP searches (Altschul et al., 1990) of the rice protein sequences against *Arabidopsis* sequences with a cutoff expect score of 1e-20. When a rice protein had more than one *Arabidopsis* protein hit, the pair having the most significant expect score was retained. In all, 4,447 homologous pairs were identified.

After the homologous protein sequences had been identified, the corresponding nucleotide sequences were scored for nucleotide content. In this study, we ranked the rice homologs by their GC content. We then compared the group of 1000 rice genes with the

highest GC content (the "High G+C" class) to their homologs in the *Arabidopsis* genome. We also performed a parallel comparison between the group of 1000 rice genes having the lowest GC content (the "Low G+C" class) and their homologs.

4.3.3 Identifying amino acids for GC-rich and AT-rich codons

In the manner introduced by Foster, Jermiin and Hickey (1997), we partitioned the universal codon table (Fig. 4.1 left) into three groups: codons that were GC-rich at the first two codon positions; those that were AU(AT)-rich at the first two codon positions; and unbiased codons (Fig. 4.1 right). The GC-rich codons encode glycine, alanine, arginine and proline (GARP). The AT-rich codons encode phenylalanine, tyrosine, methionine, isoleucine, asparagines and lysine (FYMINK). The unbiased codons fill two quadrants of the rearranged codon table and they encoded serine (S), threonine (T), cysteine (C), tryptophan (W) and valine (V), leucine (L), glutamic acid (E), aspartic acid (D), histidine (H) and glutamine (Q).

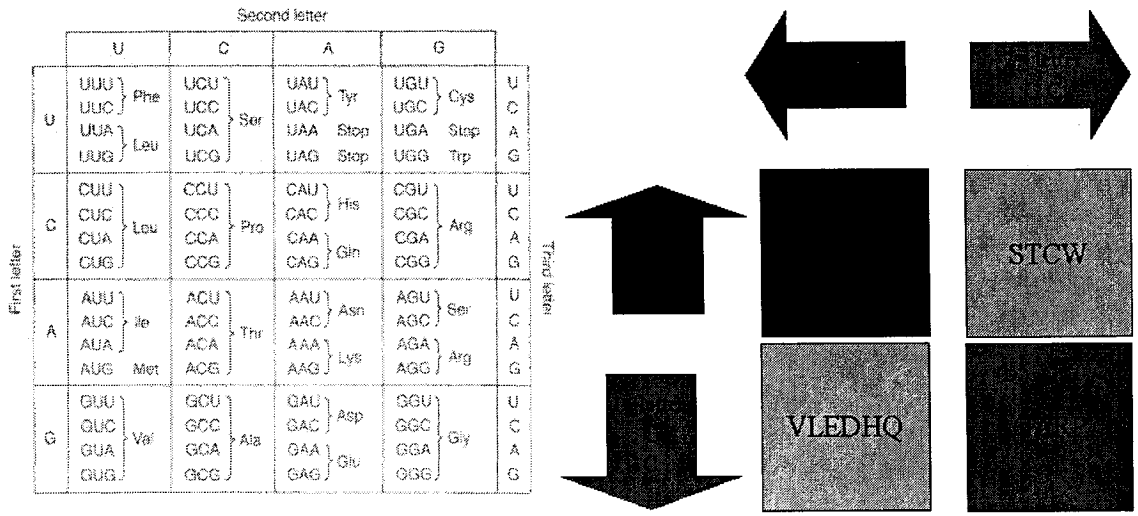


Fig. 4.1 According to GC composition at first two codon positions, the universal codon table (left panel) is partitioned into AU (AT)-rich codons (0% GC at the first 2 positions), GC-rich codons (100% GC) and the other unbiased codons (50% GC). The corresponding amino acids for the codons are shown in the four rectangle boxes, represented as single letters (right panel). Adapted from Foster, Jermiin & Hickey, 1997.

4.4 Results

4.4.1 Compositional distribution of rice and *Arabidopsis* homologous genes

First, we confirmed previous reports that the genomes of monocotyledonous plants, including rice, have elevated GC contents (Carels et al., 1998; Sasaki et al., 2002; Wong et al., 2002) and that there is a wide variation in GC content among rice genes (Carels and Bernardi, 2000; Yu et al., 2002). *Arabidopsis* genes have relatively low GC contents and they form a unimodal distribution, with a mean GC content of about 44%. Rice genes, on the other hand show a much broader, multimodal distribution (Fig. 4.2A). One possible interpretation of these results is that the rice genome contains a unique set of genes that are characterized by a higher GC content than the set of genes that is shared between the two genomes. To investigate this possibility, we repeated the same analysis based on 4,447 pairs of homologous genes that we identified using BLAST searches (see Materials and methods). As can be seen in Figure 4.2B, the same trends are seen in this subset of homologous genes. This indicates that the differences in nucleotide content are not simply due to differences in gene content between the two genomes.

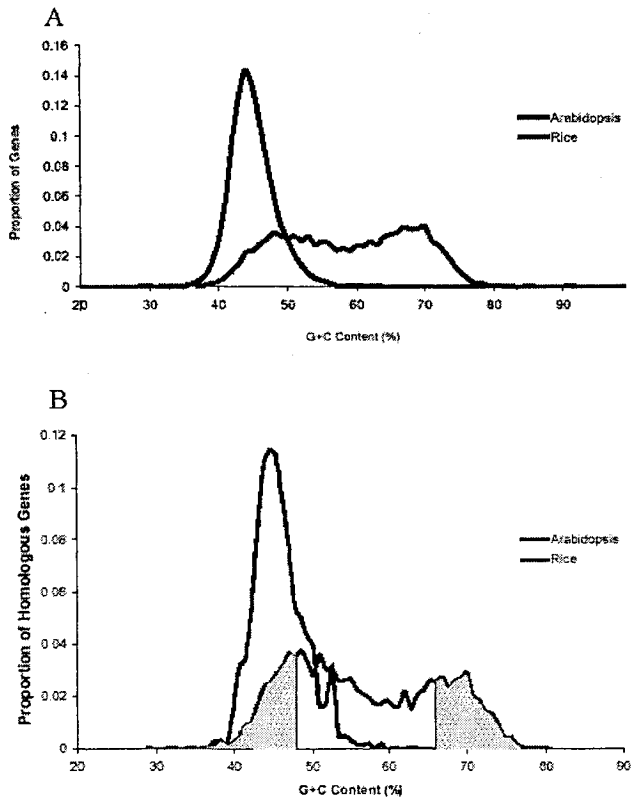


Fig. 4.2 Distribution of GC contents among rice and *Arabidopsis* genes. (A) 7886 rice genes and 25625 Arabidopsis genes. (B) 4447 homologous gene pairs of rice and *Arabidopsis* homologs. The two shaded areas in (B) mark rice homologous genes of 1000 lowest GC and 1000 highest GC contents (<47.74% and >65.91%), respectively.

For our subsequent analyses, we compared the rice genes from the two ends of the distribution shown in Fig. 4.2B to their *Arabidopsis* homologs. Specifically, we chose 1000 rice genes with the lowest GC contents as one set ("Low G+C" rice genes) and those with the highest GC content as the other set ("High G+C" rice genes). Each of these sets represents approximately one quarter of the total set of homologous sequence pairs. The "Low G+C" rice genes have nucleotide contents that lie within the distribution of their *Arabidopsis* homologs, whereas the "High G+C" genes lie well outside the *Arabidopsis* range. The average nucleotide contents of these gene sets are shown in Table 4.1. The average value for the Low GC rice genes is very close to that of their *Arabidopsis* homologs (Table 4.1c), whereas the High GC rice genes have diverged greatly from their homologs. This is especially evident at the third positions of codons where the rice genes have a GC

Table 4.1 Average nucleotide contents of homologous genes in rice and *Arabidopsis* (expressed as percentages of G+C).

	Codon Position			
	First	Second	Third	Average
(a) All Homologous pairs (n = 4447)				
Rice	58.1	44.7	66.4	56.4
Arabidopsis	51.5	41.0	43.9	45.5
(b) High GC Genes* (n = 1000)				
Rice	65.4	51.6	91.8	69.6
Arabidopsis	51.7	43.5	46.9	47.3
(c) Low GC Genes* (n = 1000)				
Rice	51.6	39.1	43.3	44.7
Arabidopsis	50.8	38.9	41.0	43.5

* "High G+C" and "Low G+C" refers to the nucleotide content of the rice genes only, not to their *Arabidopsis* homologs.
This Table shows the GC content of each codon position (along with the average value for all three codon positions). The table contains three parts: (a) Data for all homologs; (b) Data for the 1000 High GC rice genes and their *Arabidopsis* homologs; (c) Data for the 1000 Low GC rice genes and their *Arabidopsis* homologs.

content that is almost twice as large as that of their homologs (91.8% versus 46.9%, see Table 4.1b). It is interesting to note that, despite the large differences in average GC content between the two groups of rice genes, the *Arabidopsis* homologs of all of these groups have relatively constant GC contents (Table 4.1). This is consistent with the unimodal distribution of nucleotide contents among all *Arabidopsis* genes (Fig. 4.2).

4.4.2 Amino acid substitutions between rice and *Arabidopsis* homologs

The main focus of our study was to investigate the degree to which changes in nucleotide composition could influence the evolution of the encoded proteins. In particular, we asked if different rice proteins might follow different evolutionary trajectories depending on their evolving nucleotide content since the divergence of rice and *Arabidopsis*. A previous analysis of many prokaryotic genomes (Singer and Hickey, 2000) showed that proteins encoded by GC-rich sequences are characterized by increased levels of the amino acids G, A, R, and P. These proteins show a corresponding decrease of amino acids encoded by AT-rich codons - namely, F, Y, M, I, N, K. In this study, we compared the amino acid contents of proteins encoded by High GC rice genes, with their *Arabidopsis* homologs. We found that the rice genes do indeed show a highly significant increase in the level of GARP amino acids, and a corresponding decrease in FYMINK amino acids. In contrast to this, the control set of Low GC rice genes encode proteins that have amino acid contents very similar to their *Arabidopsis* homologs (data not shown).

In addition to showing simple differences in amino acid compositions between the homologous protein sequences, we wanted to investigate the patterns of amino acid substitution during the course of their evolutionary divergence. To do this, we aligned the homologous sequence pairs, and we then concatenated these alignments. The aligned sites can be classified as invariant (where the same amino acid appears in the rice and *Arabidopsis* sequences) or variant (where there is a difference between the two sequences). Since it is only these latter sites that contain information about sequence divergence, we re-calculated the amino acid frequencies for these sites only. The results for the High GC rice genes and their *Arabidopsis* homologs are shown in Fig. 4.3A. In this case, there is a two fold increase in the proportion of GARP amino acids in the rice sequences, and an even greater proportional decrease in FYMINK amino acids. Not only is there a large average difference

between the two concatenated sequences, but there is a consistent trend seen among individual homologous gene pairs. For instance, 971 out of the 1000 High GC rice genes have higher GARP levels than their *Arabidopsis* homologs and this trend is highly significant ($p << 0.00001$ in a one-tailed paired-sample *t*-test). There are also consistent differences for all of the individual amino acids, within both the GARP and FYMINK groups of amino acids (Fig. 4.3B). Some of these frequency changes for individual amino acids are quite dramatic. For instance, the rice genes have a three-fold increase in the

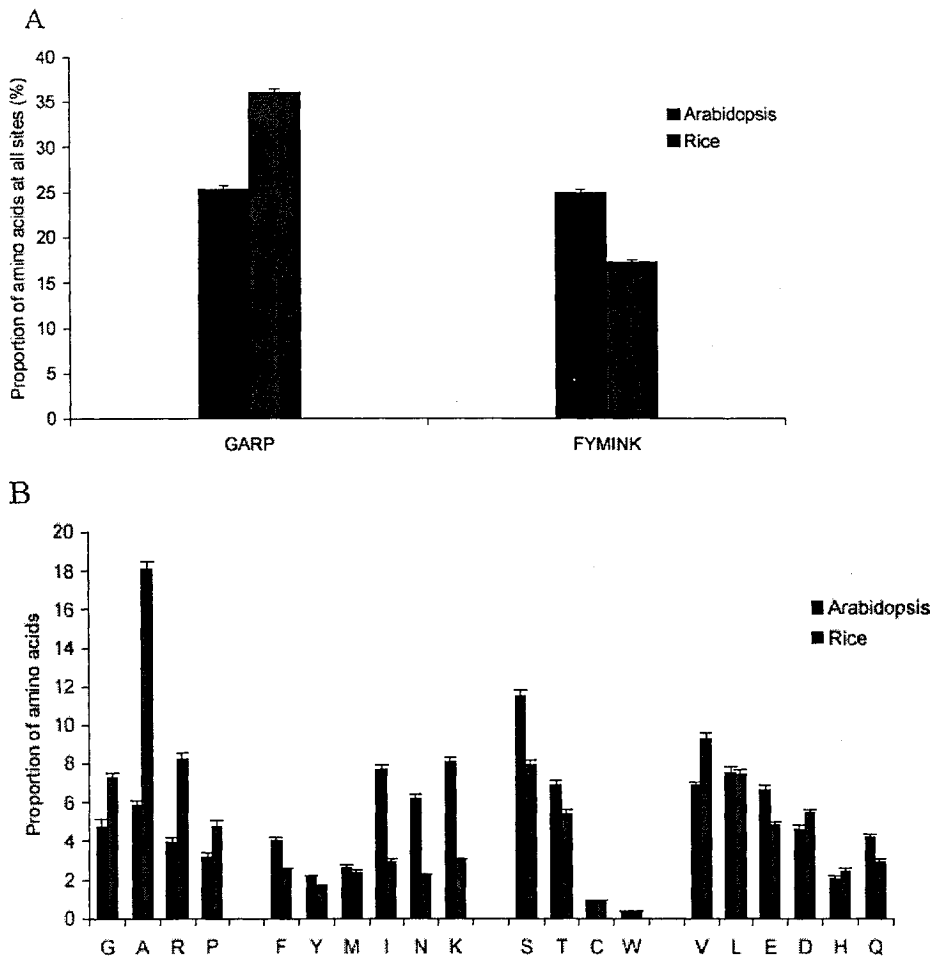


Fig. 4.3 Amino acid content of High GC rice and homologous *Arabidopsis* protein sequences. (A) The content of G,A,R,P and F,Y,M,I,N,K amino acids (expressed as percentages) for High GC rice genes and their *Arabidopsis* homologs (1000 genes each). These data are based on variant sites only in the aligned homologous sequences. (B) Proportions of individual amino acids at variant sites (expressed as numbers per 100 variant sites) plotted for the High GC rice genes and their homologs from *Arabidopsis*. The values for the rice genes are shown in red; the values for the *Arabidopsis* homologs are shown in blue. The error bars represent the 99% confidence intervals.

proportion of alanine (A) at the variant sites and a two-fold increase in arginine (R). They show a correspondingly large (more than two-fold) decreases in isoleucine (I), asparagine (N) and lysine (K). The differences in amino acid composition are highly significant ($p << 0.001$) for all but three of the twenty pair-wise comparisons. The exceptions are cysteine (C), tryptophan (W) and leucine (L).

In contrast to these large differences in the amino acid composition at variable sites of the High GC rice genes encoded proteins and their *Arabidopsis* homologs, the variable sites of the Low GC rice genes encode proteins have amino acid contents very similar to that of their *Arabidopsis* homologs (Fig. 4.4).

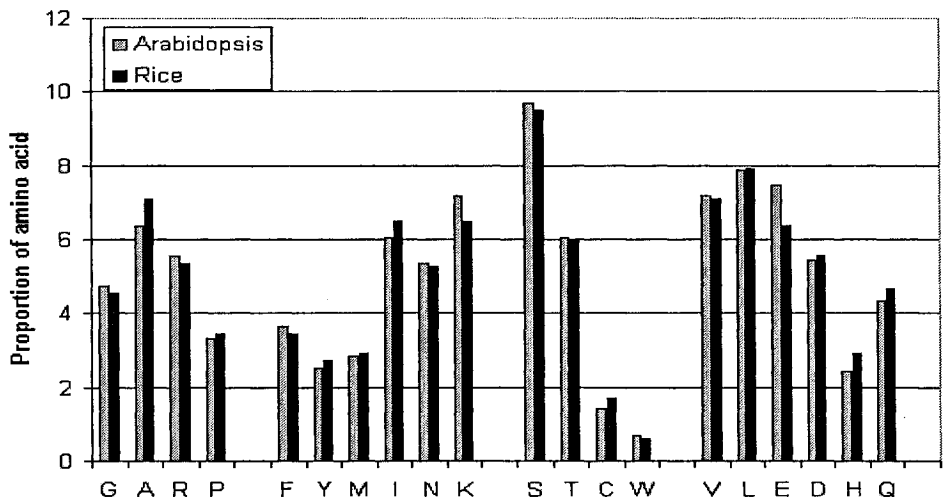


Fig. 4.4 Average amino acid contents of Low GC rice and *Arabidopsis* homologous protein sequences at variant sites (expressed as numbers per 100 variant sites). The values for the rice genes are shown in red; the values for the *Arabidopsis* homologs are shown in blue.

The fact that there are large differences in the proportions of certain amino acids at the variable sites in aligned amino acid sequences indicates that the pattern of amino acid substitution between the rice High GC genes and their *Arabidopsis* homologs is highly asymmetric. This can be seen more clearly when we construct an amino acid exchange matrix for the aligned sequences. Such a matrix is shown in Figure 4.5. In this matrix, the rows represent the rice sequence data and the columns represent the *Arabidopsis* data; the diagonal represents invariant sites. For example, there are 127 sites per 10,000 (shown in green on the matrix) where a lysine (K) in the *Arabidopsis* sequence is aligned with an

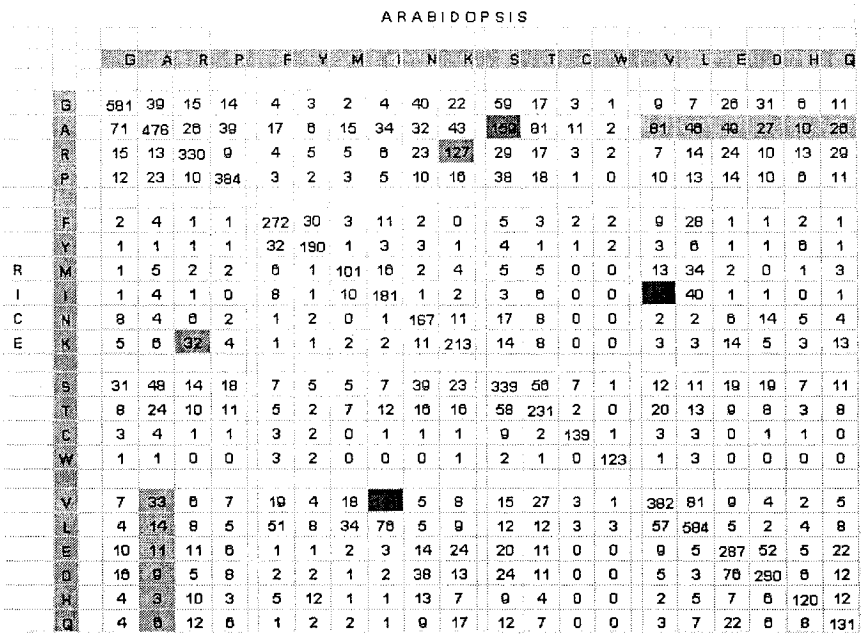


Fig. 4.5 Amino acid exchange matrix. This figure shows the pattern of sequence divergence between proteins of the High GC rice genes and their *Arabidopsis* homologs. Homologous sequences were aligned and the numbers of sequence mismatches were scored. Values in the matrix were scaled to represent the number of amino acid mismatches per 10,000 sites. Highlighted areas are discussed in the text.

arginine (R) in the rice sequence. In contrast, we see only 32 sites (also shown in green) where an arginine in the *Arabidopsis* sequence has been aligned with a lysine in the rice sequence. In other words, the great majority of arginine-lysine mismatches between the aligned sequences contain an arginine in the rice sequence and a lysine in the *Arabidopsis* sequence. Since arginine and lysine are biochemically similar amino acids, this allows the rice genes to increase their GC content while maintaining their biochemical function.

The trend illustrated above by the arginine and lysine sequence mismatches extends to the entire group of GARP and FYMINK amino acids between the two sets of homologs (see yellow-shaded quadrants in Fig. 4.5). For instance, out of a total of 526 mismatches between GARP and FYMINK amino acids, the rice sequence is overwhelmingly more likely (431 out of 526 times) to possess an amino acid from the GARP group. This represents a more than four-fold asymmetry in the pattern of amino acid substitution. Because we have a large data set of aligned homologous sequences, we are able to evaluate the rate of exchange between all 400 amino acid combinations. As expected, the rates of exchange between biochemically similar amino acids, such as isoleucine and valine, are high. What is

noteworthy, however, is that there is a strong asymmetry in these exchanges; isoleucine is found in the *Arabidopsis* sequence and valine is found in the rice sequence 144 times out of a total of 189 mismatches (see Fig. 4.5). This asymmetry that can be seen throughout the exchange matrix is indicative of significant, non-random evolutionary changes in the rice proteins as a result of the mutational bias at the DNA level.

It should be noted that it takes two nucleotide substitutions (at both the first and second codon positions) to achieve an amino acid exchange between the GARP and FYMINK amino acids. This means that, during a period of increasing GC content in the rice sequences, many of the substitutions involve "intermediate" amino acids, i.e., amino acids that are encoded by codons with intermediate nucleotide content. For instance, as the rice genes become more GC rich, they are expected to gain GARP amino acids by single nucleotide substitutions from the pool of codons with intermediate nucleotide content. At the same time, they will lose FYMINK amino acids because of mutations that change the latter codons into codons of intermediate nucleotide content. Thus, these intermediate codons act as the "flow- through" from one extreme to the other. This effect is also illustrated in Fig. 4.5. For instance, we see that the source of the huge increase in alanine (A) in the rice sequences is not primarily due to direct substitution from the FYMINK group, but from other amino acids, such as serine (S) (shown in blue). Likewise, the greatest single loss of isoleucine (I) in the rice sequences is to the G+C-intermediate valine (V, shown in red). More generally, we can see that exchanges between alanine in the rice sequences and the intermediate group of V, L, E, D, H, Q (shown in orange in Fig. 4.5) results in a net increase of 163 alanines (239 - 76 = 163). We used Fisher's exact test to calculate the significance of these asymmetries. All of the differences mentioned above are highly statistically significant ($p < 0.001$) and, in fact, the gain of Alanine from all other amino acids is significant ($p < 0.05$) except for Tyrosine (Y), Cysteine (C) and Tryptophan (W). In other words, Alanine becomes a "sink" in the High GC rice genes.

In contrast to these findings, when we constructed a parallel exchange matrix for the low GC rice gene encoded proteins and their *Arabidopsis* homologs (Fig. 4.6), we found no evidence of asymmetry in the patterns of amino acid substitution, which is consistent with their very similar amino acid compositions (Fig. 4.4). Thus the asymmetric pattern of protein evolution is correlated with the changes in nucleotide content among the rice genes.

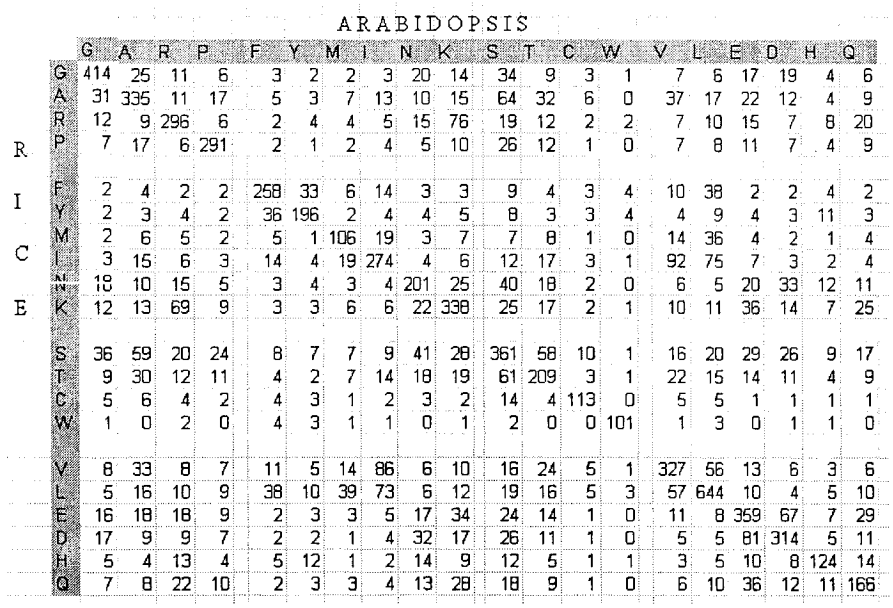


Fig. 4.6 Amino acid exchange matrix. This figure shows the pattern of sequence divergence between proteins of the Low GC rice genes and their *Arabidopsis* homologs. Homologous sequences were aligned and the numbers of sequence mismatches were scored. Values in the matrix were scaled to represent the number of amino acid mismatches per 10,000 sites.

4.4.3 Possible sources of compositional bias in rice genes and their encoded proteins

Although our primary purpose was to explore the effects of mutational bias on the patterns of protein evolution, we also wished to infer the causes of this variation in nucleotide content between rice genes. First, we wished to reconcile the reports of Carels and Bernardi (2000) who state that there are two classes of genes in plants (one class being G+C-rich) and of Wong et al., (2002) who find there is a gradient of GC content along individual rice genes. Figure 4.7 shows a plot of average GC content in a sliding window of 51 nucleotides for 4447 rice high GC and low GC genes and their *Abaridopsis* homologs. Both the rice high and low GC genes have gradients at the 5' ends, whereas gradients in the *Arabidopsis* homologs, if any, are very weak. Between the two sets of rice genes it appears that the Low GC genes has increased levels of GC at their 5' ends only, while the High GC genes have a more extended gradient. This difference should be reflected in the difference in the amino acid bias of encoded proteins, especially for the amino acids of GC-rich and AT-rich codons. Figure 4.8 is a plot of the amino acid composition (the sum of GC-rich G, A, R, P amino acids and the sum of AT-rich F, Y, M, I, N, K amino acids) along the protein sequence

positions in the High GC and Low GC rice genes encoded proteins. These illustrate that all rice genes tend to have especially elevated levels of GC-rich codons (encoding the G, A, R,

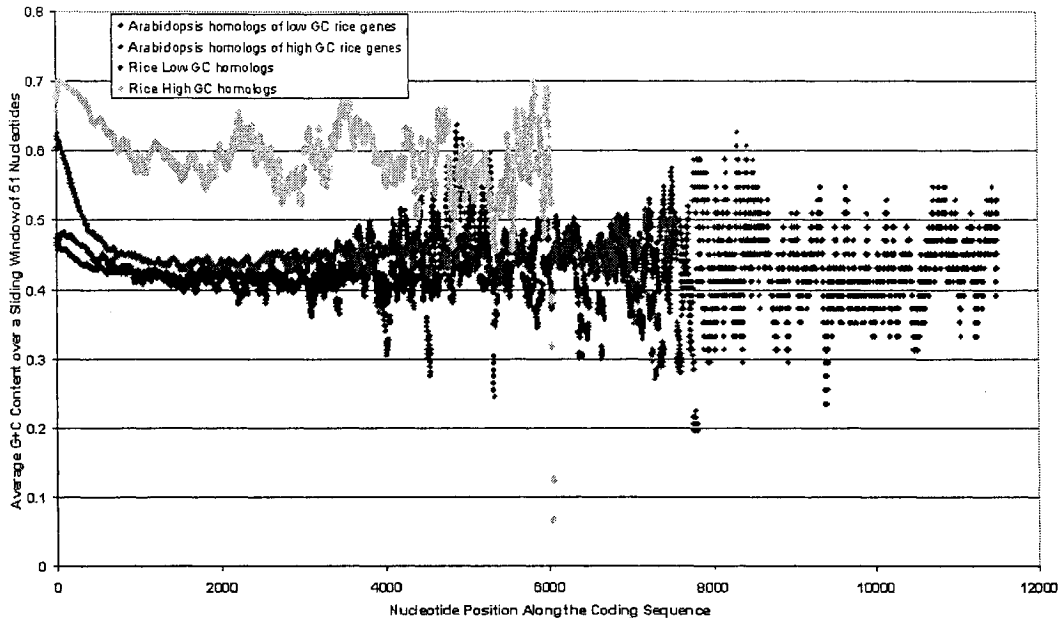


Fig. 4.7 GC content plots along the coding sequence for rice high GC, low GC genes (4447 genes in total) and their *Arabidopsis* homologs (4447 genes in total). A sliding window of 51 nucleotides was used to score the frequency of GC along the sequence positions.

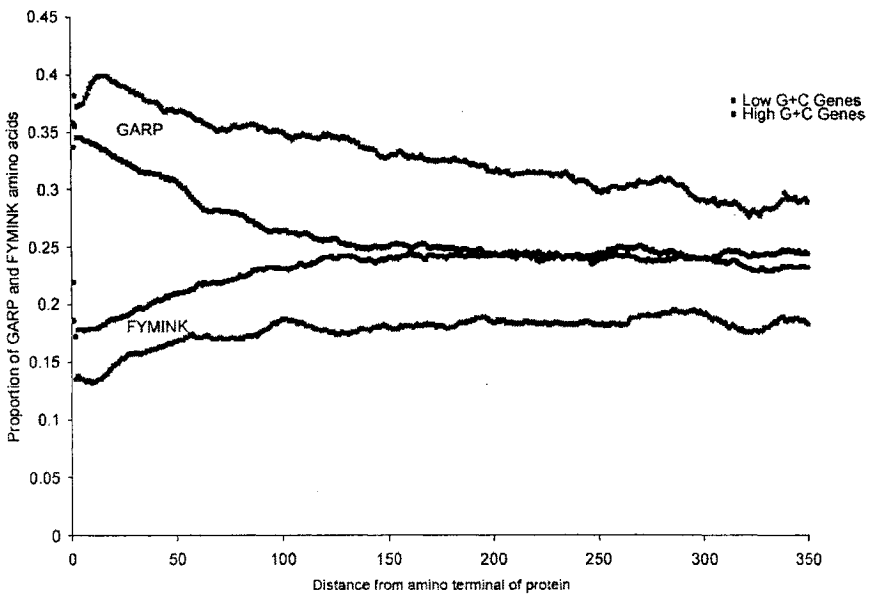


Fig. 4.8 The degree of mutational bias correlates with position within the gene. There is a gradient in the frequency of GARP (top) and FYMINK (bottom) amino acids along the encoded proteins. A sliding window of 17 amino acids was used to score the frequency of GARP and FYMINK along the first 350 amino acid positions.

P amino acids) at their 5' ends, but that the High GC class is characterized by a tendency to have this elevated level extend over the entire coding sequence, which agrees with the GC gradient plot (Fig. 4.7). In summary, we found that the differences in nucleotide composition between rice genes are due to a combination of a gradient along the gene length (as noted by Wong et al., 2002) and an overall average difference between the genes (as noted by Carels & Bernardi, 2000). Neither the compositional gradient along the coding sequence length, nor the bi-modal distribution of nucleotide composition among genes is seen in the *Arabidopsis* genome.

The existence of a compositional gradient along the coding sequence of individual rice genes suggests that the forces acting to increase the GC level of rice genes since their divergence from their angiosperm ancestors is somehow linked to the orientation of gene transcription. This led us to wonder if there might be some strand asymmetry in the pattern of bias, as seen in some prokaryotic and organelle genomes (Lobry, 1996; Morton, 1999; Tillier and Collins, 2000). We investigated this by calculating the frequencies of individual nucleotides, rather than the sum of G plus C. The result (not shown) revealed no indication of strand asymmetry: the levels of both G and C were equally elevated in the GC rich genes and the levels of A and T were equally reduced. To our surprise, however, we did find a very strong link between gene length and GC content (see Fig. 4.9). Specifically, the High GC rice genes are much shorter, on average, than the low GC genes. One's first thought is that the increased nucleotide bias might somehow act to decrease gene length. This unlikely

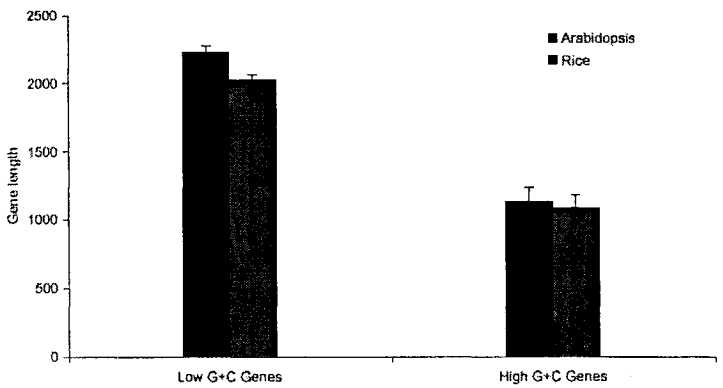


Fig. 4.9 The relationship between coding sequence length and GC content. On the right, the average lengths and 99% confidence intervals are shown for the High GC rice genes and their *Arabidopsis* homologs. The corresponding data for the Low GC rice genes and their *Arabidopsis* homologs are shown on the left.

scenario can quickly be discounted, however, by noting that the difference in gene length is equally impressive for the *Arabidopsis* homologs, all of which have virtually identical nucleotide contents. The most parsimonious explanation is that shorter genes are more susceptible to whatever mutational forces are causing some of the rice genes to become GC rich.

4.4.4 Mutational bias affects protein sequence similarity

Having demonstrated that amino acid substitutions in the proteins encoded by G+C-rich rice genes are significantly affected by nucleotide bias, we asked how this might affect computer-based estimates of sequence similarity between rice and arabidopsis proteins. In Figure 4.10A, we show the distribution of BLAST scores for the High GC and Low GC rice genes when compared to their arabidopsis homologs. To the left of the Figure, we can see that there is preponderance of Low GC genes among the highest scoring comparisons. This changes gradually, however, and for the lower scores, at the right of Figure 4.10A, we see that there is a majority of High GC sequences among the lower scoring classes. These results show that there is a wide range of sequence divergences among both the Low GC and the High GC encoded proteins, but there is also a tendency for the higher GC content to lower the BLAST score at all levels of divergence. We have summarized these results using all of the rice genes that have homologs in the *Arabidopsis* protein database (not only the High GC and Low GC sets). The results are shown in Figure 4.10B. It is clear from the Figure that there is indeed a significant effect of nucleotide bias on these BLASTP scores, and that it is in the predicted direction. For instance, those rice genes with elevated GC content, and with parallel increases in GARP amino acids in their encoded proteins, have decreased levels of similarity with their arabidopsis homologs. This means that the mutational bias toward higher GC not only affects the amino acid content of the encoded proteins, but it also accelerates the rate of sequence divergence between those proteins and their *Arabidopsis* homologs.

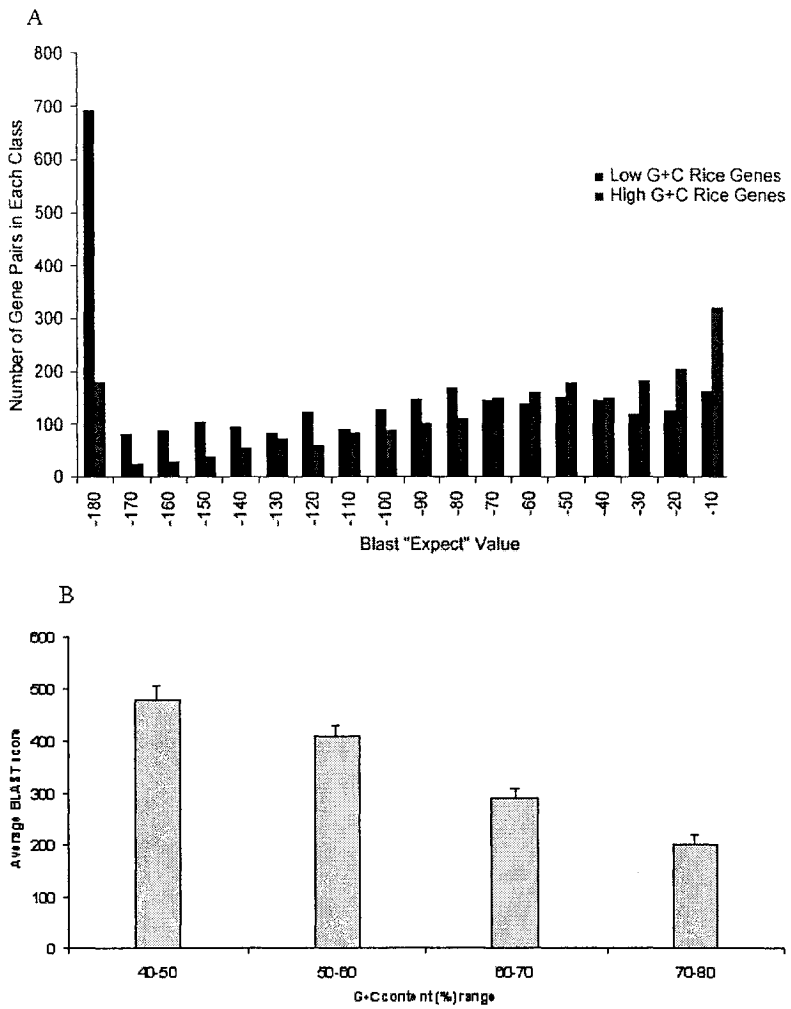


Fig. 4.10 The correlation between sequence divergence and the GC content of rice homologs. (A) BLASTP searches of the rice protein sequences were performed against the *Arabidopsis* protein sequences. The results (4447 pairs of homologous sequences) were sorted by the "expect" value, shown on the X axis. For each level of Blast e-values, the number of High GC and Low GC rice genes were scored. Note the excess of Low GC genes (shown in blue) at the left of the Figure and the excess of High GC at the right. (B) Average BLAST similarity scores among the four GC content ranges of the 4447 rice homologous sequences. Standard error bars are shown.

4.5 Discussion

Our results show a clear correlation between the variations in nucleotide composition of different rice genes and the evolutionary changes in the amino acid composition of their encoded proteins. Such a correlation could reflect either a primary effect at the level of nucleotide bias - that produces a secondary effect at the protein level or, alternatively, it

could be due to selection for amino acid content at the protein level. The first indication that mutational bias at the nucleotide level is, indeed, the primary cause comes from the observation that the differences in GC content are greatest at the third position of codons (see Table 4.1b). If the changes in average nucleotide content were due to a primary effect at the amino acid level, we would expect that the greatest change would be at the first and second positions of codons. A related method for distinguishing between nucleotide-level and protein-level effects is to compare the calculated rates of synonymous and non-synonymous nucleotide substitutions. We used the method of Yang and Nielsen (2000) to calculate these rates for the two groups of rice genes (High GC and Low G+C) compared to their *Arabidopsis* homologs. If the nucleotide composition of the High GC rice genes is affected primarily by selection at the protein level, we should see elevated rates of non-synonymous changes. If, on the other hand, the primary effect is at the nucleotide level, we should see an elevation in the synonymous substitution rate. The results show very clearly that the increase in substitution rate happens at the synonymous sites, where the average rates (mean $\pm$ standard error) for the High GC genes and the Low GC genes are 3.87 ± 0.02 and 3.50 ± 0.04, respectively. The non-synonymous substitution rate remains relatively constant between the two sets of genes ($dN = 0.5 \pm 0.009$ in both classes) (Fig. 4.11). Figure 4.12 shows a histogram of the relative ratio of substitutions (dN/dS) for the Low GC and High GC rice genes compared with Arabidopsis homologs. Except for seven High GC rice genes, which encode glycine-rich cell wall structural protein, proline-rich extension-like family protein, arabinogalactan protein, leucine-rich repeat family protein, protoporphyrinogen oxidase and two unknown hypothetical proteins, all other 993 High GC genes and all 1000 Low GC genes have a dN/dS ratio less than 1 (Fig. 4.12). This indicates positive selection, if any, is extremely rare in the two rice gene sets compared with the *Arabidopsis* homologs. All these results point to mutational bias at the nucleotide level, rather than functional selection at the protein level.

Our results indicate that the High GC genes are undergoing accelerated rates of sequence divergence from their *Arabidopsis* homologs (see Fig. 4.10). In other words, different rice genes tell different "evolutionary stories" depending on the degree to which they are affected by mutational bias. Previous studies of mitochondrial genomes have shown that such biases may lead to erroneous phylogenetic reconstructions (Foster and Hickey,

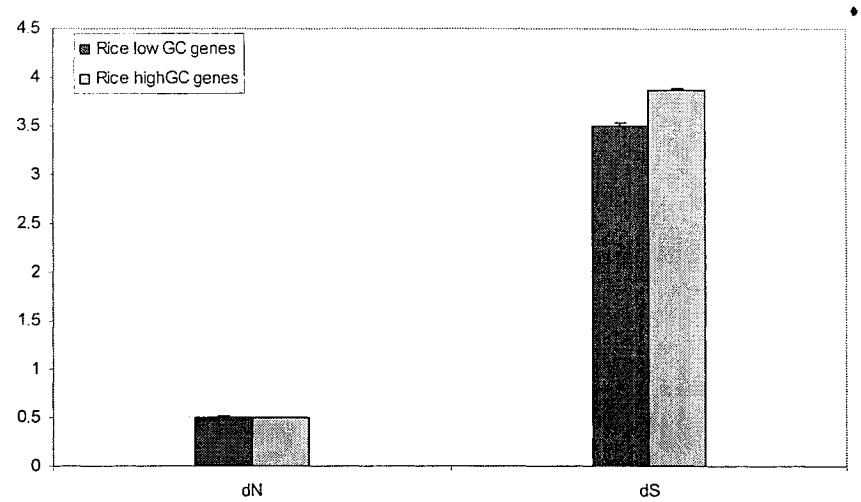
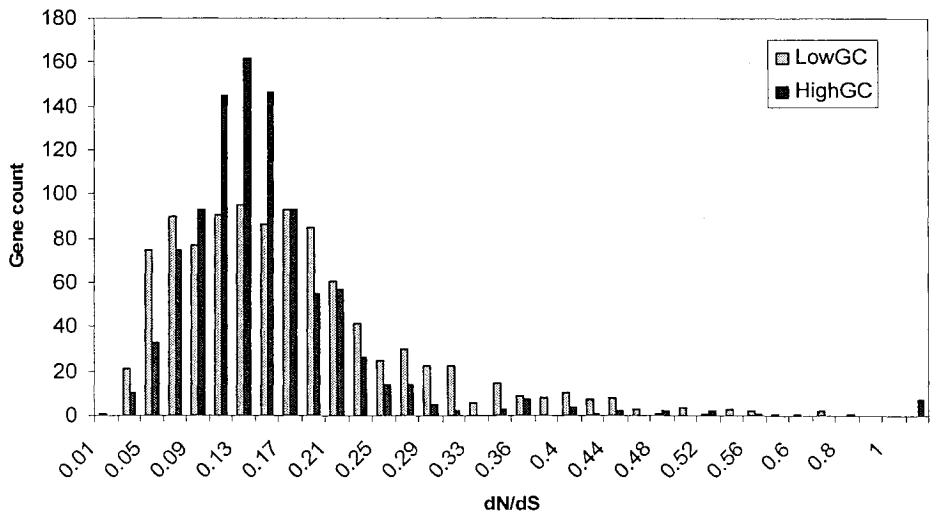


Fig. 4.11 dN and dS in rice Low GC genes and High GC genes compared to *Arabidopsis* homologs. For each panel of the two bars value for the rice low GC genes is on the left and that for the high GC genes is on the right. Average dN and dS and standard error (error bar) for rice High GC and Low GC genes. dS are the average of all dS values that were calculated from the two rice gene sets using the PAML program (Yang & Nielsen, 2000), some of which contain values of 99, which is the case when PAML cannot calculate the dS due to the two sequences are too divergent and they have reached substitution saturation. The dS cannot be correctly calculated so PAML arbitrarily assigned a high value (99). The Low GC gene set has one dS value of 99 while the High GC set has 39 dS value of 99. The averages of the dS shown here were calculated when the value(s) of 99 were removed.

Fig. 4.12 (below) Histogram of dN/dS distribution in the rice low GC and high GC gene sets compared to Arabidopsis homologs. For each set of the two bars, value for rice low GC genes is on the left and that for high GC genes is on the right.



1999). This is because most methods of phylogenetic reconstruction assume that all of the sequences share a common pattern of nucleotide substitution since their divergence from a common ancestor (Tamura & Kumar, 2002). These authors have introduced a correction to reduce the effects of heterogeneity in substitution pattern on the estimation of evolutionary distances between sequences. We tested the ability of this method to correct for the effects of mutational bias in the High GC rice genes. In Figure 4.13, we show both uncorrected and corrected evolutionary distances, calculated by using Tamura & Nei (1993) method and Tamura & Kumar (2002) method, respectively. As can be seen from the Figure, the Tamura-Nei distance between the High GC rice genes and their *Arabidopsis* homologs was greatly increased compared to that of the Low GC rice genes and their homologs despite the fact that the actual divergence time is the same for the two groups. When we used the correction of Tamura and Kumar (2002), this difference is greatly reduced, although not completely eliminated (Fig. 4.13). Historically, the recognition of nucleotide bias as a pervasive feature

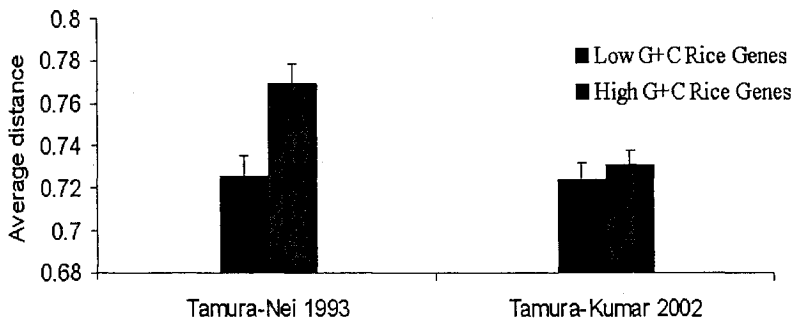


Fig. 4.13 Average evolutionary distance of Low GC and High GC rice genes and *Arabidopsis* homologs, calculated using Tamura & Nei (1993) method (uncorrected distance) and Tamura & Kumar (2002) method (corrected distances). For each panel of the two bars values for the rice low GC genes are on the left and that for the high GC genes are on the right.

of molecular evolution led to a shift of emphasis from nucleotide to amino acid sequences for phylogenetic reconstruction (Loomis and Smith, 1990; Hasegawa and Hashimoto, 1993). Subsequently it was shown, however, that this problem of nucleotide bias can affect phylogenies based on the derived amino acid sequences (Foster and Hickey, 1999). The approach of Tamura and Kumar (2002) tackles the problems of nucleotide bias and heterogeneous substitution rates directly, rather than simply trying to "hide" the problem in the derived amino acid sequences.

Our finding that shorter coding sequences have a greater tendency to increase in GC content confirms the findings of Carels and Bernardi (2000) and it is reminiscent of the finding of Duret et al. (1995) who showed that the GC content of many vertebrate genes was negatively correlated with coding sequence length. More recently, a similar trend has been noted in a survey of single-exon coding sequences in five eukaryotic species (Xia et al, 2003). It is intriguing to observe the same length correlations in both vertebrates and plants. One possible explanation for this trend is that the increased GC content is linked to the absence of introns, if one assumes that longer genes are more likely to have multiple exons. However, we tested this hypothesis by confining our analysis to single-exon genes only, and we found that the presence of introns was not the primary determining factor of nucleotide content. For instance, with this more restricted data set (single exon genes only), the difference in gene length was just as great as that shown in Figure 4.9 for all genes. Thus the length difference is maintained even in the absence of introns. Interestingly, however, we found that the High GC rice genes included relatively few multi-exon genes, especially genes with three or more exons (see Table 4.2). This suggests that the presence of multiple

Table 4.2 Exon-intron structure of rice genes and their *Arabidopsis* homologs.

	Number of Exons			
	1	2	3	4+
(a) High GC Genes* (n=1000)				
Rice	434	294	165	107
Arabidopsis	308	224	182	286
(b) Low GC Genes* (n=1000)				
Rice	260	72	165	503
Arabidopsis	159	89	57	695

Note: Two general trends can be seen: (i) rice genes tend to have fewer exons, on average, than their *Arabidopsis* homologs; and (ii) the High GC rice genes - and their *Arabidopsis* homologs - have fewer exons than the Low GC genes. These differences are highly significant ($p << 0.0001$ in a Chi Square test).

* "High G+C" and "Low G+C" refers to the nucleotide content of the rice genes only, not to their *Arabidopsis* homologs. The numbers in the table refer to the number of genes that fall into each exon class.

introns may prevent even short genes from becoming GC rich. This is supported by the observation that among the low GC rice genes, the average length of three-exon and four-exon genes is only 60% of the length of one-exon and two-exon genes. In other words, even though the coding sequences of these genes are relatively short, the presence of multiple introns may prevent them from becoming GC rich. This implies that there are selective constraints related to RNA splicing that counter the effects of mutational bias in these genes. Such a constraint would not, however, explain the fact that long, single-exon coding sequences also remain relatively immune to mutational bias. The answer may lie in the fact that RNA splicing is only one form of RNA processing. In general, it may be that longer genes encode more complex transcripts and proteins that have a greater chance of being functionally disrupted by biased mutational changes. For instance, one kind of the mutations is cytosine deamination which coverts C to U (see Chapter 1.3.3), a process catalyzed by deaminases of the group of RNA editing enzymes in plants. This "C to U" editing appears to be quite universal for both nucleus-encoded RNAs and organelle (i.e., mitochondria and chloroplasts)-encoded RNAs. The propensity of C mutating to U implies that an RNA molecule should not have too many C's. This in turn implies that long mRNAs should have a small proportion of C but short mRNAs can be allowed to have a larger proportion of C. This explains why long genes have on average less GC (X. Xia, personal communication). Shorter genes are also at risk, but they provide a smaller target for these mutations and, consequently, they are subject to lesser selective constraint and therefore can be more GC-rich.

Differences in coding sequence length might explain why some, but not all, rice gene are subject to mutational bias. However, they cannot explain the differences in nucleotide content between the two plant genomes. A possible explanation for the inter-genomic difference is the fact that the rice genome is much larger than the *Arabidopsis* genome and that it contains more genes (Sasaki et al. 2002). This would lead one to predict that gene families in rice may contain more members than in *Arabidopsis*. In fact, Sasaki et al. (2002) reported that a significant number of the genes found on rice Chromosome 1 are duplicated and arrayed in tandem. This difference in the size of gene families could affect the nucleotide content of the coding sequences because it has been shown that gene conversion between members of gene families can lead to increasing GC content of the converted

sequences (Hickey et al, 1991; Galtier, 2003). Another difference between the two genomes is that rice genes appear to have a lower average number of introns per coding sequence than do their *Arabidopsis* homologs (Carels and Bernardi, 2000 and see Table 4.2).

In summary, we have shown that mutational bias can have profound effects on the patterns of evolutionary divergence between homologous plant protein sequences. This indicates that mutational bias can be a major determinant of the patterns of protein evolution in eukaryotes. The rice genome does not, however, have a uniformly elevated GC content among its coding sequences. The result of this heterogeneity in the nucleotide content among the coding sequences is reflected in the very different amino acid compositions among the encoded proteins.

Chapter 5

Nucleotide content affects synonymous codon usage in rice genes

5.1 Abstract

We analyzed the codon usage patterns of 14005 *Oryza sativa* genes and compared a subset of these genes with 7160 homologous *Arabidopsis thaliana* genes. The nucleotide composition of third codon positions shows that some of the rice genes, especially high GC rice genes, have a preference for cytosine (C) over uracil (U) and guanine (G) over adenine (A), while the *Arabidopsis* genes and some low GC rice genes have a preference for U and A. Correspondence analysis indicated that the codon usage patterns are distributed according to the GC content of the individual genes, with high GC rice genes located at one end of the distribution, low GC rice genes and *Arabidopsis* genes (also in low GC) overlapped and located at the other end, while the intermediate GC rice genes located in the middle. The underlying codons were also distributed accordingly with codons of high GC content located at the high GC rice region and codons of high AT content at the *Arabidopsis* gene or low GC rice gene regions. These results suggest that regional compositional bias is a major factor shaping rice codon usage. But it may not be solely responsible for the variation in GC content at synonymous third codon sites (GC3s), as the correlation between GC3s and intron GC content is weaker than the correlation between GC3s and GC of coding regions. Other factors include gene length – short genes were found to have greater codon bias, and selection pressures are considered. The effective number of codons and codon adaptation index of the rice genes indicated translational selection is not as obvious as in *Arabidopsis* codon bias, which is further supported the evidence that a positive, rather than inverse, correlation, between codon bias and the rate of molecular evolution at the synonymous sites in the rice genes.

5.2 Introduction

Although the choice among synonymous codons does not alter the amino acid sequence of the encoded protein, numerous studies in the past 25 years have found that the choice of synonymous codons in most organisms is not random. Codon usage patterns vary among genes within a genome, and also among genomes. It has been proposed that there exists a fundamental dichotomy between the codon usage of unicellular and multicellular organisms (Peden, 1999). For prokaryotes and unicellular eukaryotes such as yeast, the variation was considered to be due to natural selection acting to optimize protein production. According to this proposal, highly expressed genes use codons that are complementary to abundant tRNA anticodons, whereas other genes use different sets of synonymous codons depending on their expression level. Some multicelluar organisms, such as *Drosophila melanogaster* and *Caenorhabditis elegans* also display a codon bias that appears to be caused by selection for translation efficiency. But for the majority of multicelluar organisms codon usage bias was thought to be caused by variation among chromosomal regions in DNA mutation process, for example biased mutation, resulting in nucleotide composition variation (see Chapter 1.3). Furthermore, horizontally transferred genes (or alien genes) tend to have a codon usage different from that of the host organism (Médigue et al. 1991; Lawrence and Ochman 1997; Wang et al., 2001). Therefore, analyzing codon usage will reveal information about genome evolution.

In Chapter 4 we made a comparative study of thousands of *O. sativa* (rice) and *A. theliana* homologous genes and found a strong correlation between nucleotide content and the pattern of protein sequence substitution. We concluded that nucleotide bias is the cause, rather than the consequence, of the amino acid bias. Indeed, the differences in G+C content between rice and *Arabidopsis* homologous genes are greatest at the third codon position (12.5%) and smallest at the second codon position (only 3.7%). This difference is even bigger (44.9% versus 8.1%) when rice genes of High G+C content and their *Arabidopsis* homologs are compared (see Table 4.1). These results imply that not only were the primary causes of protein sequence substitutions acting at the nucleotide level, but that we would also see a large difference in the pattern of synonymous codon usage between high G+C rice genes and their *Arabidopsis* homologs (and between the

high and low G+C of rice genes). Indeed a recent study of codon usage of 1860 rice coding sequences has suggested nucleotide mutation bias is the major factor in shaping the codon usage of rice genes (Liu et al., 2004). In this study we employed a much large rice dataset and because of the nature of GC content heterogeneity of rice genes we separated the rice genes into high GC, intermediate GC and low GC classes and compared the codon usage between the two classes and between rice genes and *Arabidopsis* homologs.

5.3 Materials and Methods

5.3.1 Coding sequence data

Rice coding sequences were derived as in Chapter 4.3.1 but with expansion of the data set. Specifically, all 15130 rice protein sequences were obtained from the Gramene database as of September 2003 (Ware et al. 2002) (ftp://www.gramene.org/pub/gramene/protein/sequence/rice_sptrembl.fa). From the protein sequence identifiers we got 2748 EMBL entries that have cross references to the proteins. From the EMBL entries we extracted 14658 coding sequences that code for proteins whose accession numbers are in the list of Gramene protein accession numbers. In order to see whether the CDS sequences are the same as that in the Gramene's, we translated the 14658 CDS to protein sequences using EMBOSS program Transeq (Rice, Longden and Bleasby, 2000) and compared the translated sequences with the Gramene sequences. Comparing the two sets, we found 14437 identical sequences and 221 different sequences. The different sequences were found to be those sequences whose CDS length are not a multiple of three, so there are 1 or 2 bases at the end of the sequences that cannot be assigned to an amino acid. In these cases Gramene just terminated the amino acid sequence before the last one or two bases, while the EMBOSS program assigned an X as terminal amino acid, causing the two sequences to be different. Thus those 221 CDS sequences are also correct.

We then removed CDS sequences that are shorter than 75 codons, which left a total of 14024 sequences. Passing the sequences to CodonW (Peden, 1999) for codon integrity tests, one sequence was found to contain internal stop codons and thus removed.

Eighteen redundant sequences were also removed, leaving 14005 sequences in the final nonredundant set of rice coding sequences.

For the *A. thaliana* coding sequences we used the file containing 25625 *Arabidopsis* coding sequences (all greater than 75 codons) that we obtained previously (see Chapter 4.3.1).

5.3.2 Identification of homologous sequences

Homologous protein pairs between *O. sativa* and *A. thaliana* were identified by performing BLASTP searches (Altschul et al., 1990) of the rice protein sequences against *Arabidopsis* sequences with a cutoff Expect value of 1e-20. When a rice protein has more than one *Arabidopsis* protein hit, the pair having the lowest Expect value was retained. In all, 7160 homologous pairs were identified.

To calculate the synonymous substitution rate, 925 homologous pairs of proteins of rice and *Arabidopsis* chromosome 4 genes (gene ID start with At4g) were picked from the BLAST result and each pair was individually re-aligned using Clustalw (Thompson, Higgins & Gibson 1994). We used a Unix C shell script that automatically ran clustalw for each of the 925 pairs of protein sequences. The corresponding DNA sequences were aligned based on the protein alignments. Then the yn00 program in the PAML package (Yang & Nielsen, 2000; http://abacus.gene.ucl.ac.uk/software/paml.html#introduction) was used to calculate dS and dN values for each pair of the aligned DNA sequences, also using a Unix C shell script. When pairs of the sequences are too divergent, the synonymous substitutions reach saturation, yn00 cannot evaluate the dS value and an arbitrary value of 99 is assigned. The value is not meaningful (Z. Yang, personal communication). 30 pairs have dS of 99 and they were excluded in the subsequent substitution rate analysis.

5.3.3 Statistical analyses

5.3.3.1 G+C content and GC disparity

GC1, GC2, GC3 and GC are the frequencies of G+C of a coding sequence at first, second, third codon positions and all three positions, respectively. They were calculated using our program cob.c (Wang et al., 2001). GC disparity is a measure of the deviation

of GC content at the three coding positions (GC1, GC2 and GC3) from the average G+C of the coding sequence (GC) (Chen & Zhang, 2003).

5.3.3.2 Codon usage indices

GC3s is the frequency of G+C at the third synonymous codon position (excluding Met, Trp and stop codons).

The relative synonymous codon usage (RSCU) (Sharp and Li, 1987) is calculated as follows:

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}} \qquad \textbf{(5.1)}$$

where X_{ij} is the frequency of occurrence of the jth codon for the ith amino acid and n_i is the number of codons for the ith amino acid (ith codon family). An RSCU value close to 1.0 indicates a lack of codon bias. RSCU values are largely independent of amino acid composition and are particularly useful in comparing codon usage among genes, or sets of genes that differ in their size and amino acid composition. The RSCU values were calculated using DAMBE (http://aix1.uottawa.ca/~xxia/software/software.htm; Xia, 2000).

Codon usage entropy (Badger, 1999; Wang et al., 2001) uses Shannon information theory (Shannon, 1948) to estimate degree of deviation from equal usage of synonymous codons. Since deviation from equal codon usage is a concept that is independent of biological knowledge, it can be defined in purely mathematical terms, based on no biological assumption (such as mutational bias and translational selection) (Suzuki, Saito & Tomita, 2004). The advantage of using Shannon entropy is that it allows a complex source of bias to be represented by a single statistic. The Shannon entropy H for M possible outcomes is given by the following formula (Shannon 1948).

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i \qquad \textbf{(5.2)}$$

where P_i is the probability of the i^{th} outcome. To apply Shannon entropy to codon usage, we calculated the average index of the 20 amino acids and the 59 sense codons (the codon usage entropy) as follows (Badger, 1999; Wang et al., 2001):

$$H' = -\sum_{a=1}^{20} f_a \sum_{ca=1}^{n_a} f_{ca} \log_2 f_{ca} \qquad \textbf{(5.3)}$$

where f_a is the frequency of codons encoding amino acid a, n_a is the number of synonymous codons for amino acid a, and f_{ca} is the frequency of codon c among synonymous codons of amino acid a. The term $-\Sigma f_{ca}\log_2 f_{ca}$ measures the codon bias for codons encoding amino acid a. H' is the average codon bias among all 20 amino acids. When the frequency of every codon is 1/64, H' = 1.76 bits. Biased codon usage will have smaller Shannon entropy than unbiased usage. The program cob.c (Wang et al., 2001) was used to compute codon entropy. Because Equation 5.3 involves logarithm calculation, entropy for some short genes may not be calculated and thus were ignored by the program.

The effective number of codons (Nc) used in a gene is a measure of synonymous codon usage bias (Wright, 1990). It is analogous to the effective number of alleles n_e used at a gene locus (Kimura & Crow, 1964). An estimate of n_e is $\hat{n}_e = 1/\hat{F}$, where $\hat{F}$ is an estimate of the expected heterozygosity at the locus. Wright (1990) defined an estimate of the homozygosity of codon usage for an amino acid a as follows:

$$\hat{F}_a = \left(n_a \sum_{i=1}^{k} p_i^{\,2} - 1 \right) \Big/ \left(n_a - 1 \right) \qquad \textbf{(5.4)}$$

where p_i is the frequency of the ith codon, k is the number of synonymous codons for the amino acid a, and n_a is the number of the amino acid. The average of the $\hat{F}$ for each r-fold redundancy class (*i.e.*, onefold, twofold, threefold, fourfold and sixfold) is then calculated as:

$$\hat{F}_r = \frac{1}{n_{RC}} \sum_{a \in RC} \hat{F}_a \qquad \textbf{(5.5)}$$

where n_{RC} is the number of amino acids for the RC redundancy class. Finally, $\hat{N}_c$ is calculated as follows:

$$\hat{N}_c = 2 + (9/\hat{F}_2) + (1/\hat{F}_3) + (5/\hat{F}_4) + (3/\hat{F}_6) \quad \textbf{(5.6)}$$

The Nc takes a value between 20, when only one synonymous codon is used for each amino acid, and 61, when all codons are uniformly used. Lower Nc values indicate stronger bias. Since Nc is constrained by G+C content of the gene, it is often plotted against GC3s of the gene to investigate patterns of codon usage (Wright, 1990). In this plot a reference line, labeled as GC(ref) is computed, which indicates the expected position of genes whose codon usage is only determined by variation in GC3s. The *Nc*(GC ref) is calculated as follows (F. Wright, personal communication):

$$N_c(GCref) = 2 + GC3s + 29/(GC3s^2 + (1 - GC3s)^2) \quad \textbf{(5.7)}$$

The codon adaptation index (CAI) is a measure of the similarity of the codon usage of a particular gene to the codon usage pattern of highly expressed genes in the same genome (Sharp & Li, 1987a). For each codon in a codon family a weight w_{ij} is calculated as the relative adaptiveness of that codon:

$$w_{ij} = \frac{RSCU_{ij}}{RSCU_{i\max}} = \frac{X_{ij}}{X_{i\max}} \quad \textbf{(5.8)}$$

Where $RSCU_{ij}$ and $RSCU_{imax}$ are the RSCU of the *j*th codon and the maximum RSCU in the *i*th codon family, respectively. The denominators in RSCU cancel out so the weight is simply the ratio of the usage of each codon (X_{ij}) to that of the most abundant codon (X_{imax}) for the same amino acid. The CAI of a gene is then defined as the geometric means of the weights:

$$CAI = \left(\prod_{k=1}^{L} w_k \right)^{1/L} \quad \textbf{(5.9)}$$

where L is the number of codons in the gene and w_k is the weight of the *k*th codon in the gene sequence. Genes with CAI value close to 1 are those genes whose codon usage

matches that of the highly expressed reference set of genes. This is the classical way to calculate CAI in prokaryotes (Sharp & Li, 1987). In many eukaryotic organisms codon usage is known affected by gene length. To take into account this effect Carbone, Zinovyev and Kepes (2003) proposed the following equation to modify the weight w_{ij} in computing CAI.

$$w_{ij} = \frac{|S^i|}{|S|} \times \frac{X_{ij}}{X_{i\max}} \qquad \textbf{(5.10)}$$

where S^i is the set of coding sequences in S that contain at least one occurrence of codon i, and $|S^i|$, $|S|$ denotes the number of coding sequences in S^i and S. The CAI is then calculated as in Equation 5.9 but using the modified weight.

The calculation of the CAI depends on the identification of a predefined reference set, which contains highly expressed genes that display a dominant translational bias. An incorrect definition of the dominant codon usage bias would result in the CAI predicting an incorrect gene expression level (Grocock & Sharp, 2002). In this study we used a computer algorithm that can automatically extract a reference set and detect the most dominant codon bias, regardless of whether this bias is translational or not (Carbone, Zinovyev & Kepes, 2003). CAI Java, available from http://www.ihes.fr/~materials/description.html, implements the algorithm and uses both the original and the revised methods to calculate CAI. It was applied to the 14005 rice genes and a reference set was obtained after 20 iterations when the program reached convergence, i.e., the CAI values do not change over the last iteration.

5.3.3.3 Correspondence analysis

Following the pioneering work of Grantham et al. (1980), many multivariate statistical methods have been applied to the analysis of codon usage of various organisms. In essence, these procedures consider gene sequences as points in multidimensional space, with each dimension representing the frequency of a given codon (Wang et al., 2001). One of commonly used methods is the correspondence analysis, which creates a series of orthogonal axes to identify trends that explain the data variation, with each subsequent axis explaining a decreasing amount of the variation (Greenacre, 1984). The method, as

implemented in CodonW version 1.4 (Paden, 1999; http://www.molbiol.ox.ac.uk/cu/), was used to explore the variation of the 59 RSCU values for each of the sense codon among the compared genes. Correspondence analysis assigns ordination for each gene and codon on these axes, and the ordination of the genes and codons can be superimposed. Since the first two axes capture a larger fraction of the variance of the data than any of the other axes, genes and codons were plotted on these two axes only.

5.4 Results

5.4.1 G+C content distribution and G+C disparity

Figure 5.1 shows a bimodal distribution of G+C content in the 14005 rice genes, which is consistent with our previous data with 7886 rice genes (Chapter 4.4.1). Based on this feature we separated the rice genes as high G+C set (>65% G+C; 4213 genes), intermediate G+C set (between 50 and 65% G+C, 5892 genes) and low G+C set (≤ 50% G+C; 3900 genes). The average GC content of the high G+C set (69.7 ± 3.13 for mean ± S.D.) and that of the low G+C set (45.6 ± 3.04) are very close to the average GC content of the 1000 High GC rice genes and that of the 1000 Low GC rice genes defined in Chapter 4 (Table 4.1), respectively. The average GC content of the intermediate G+C set is 57.5 ± 4.5.

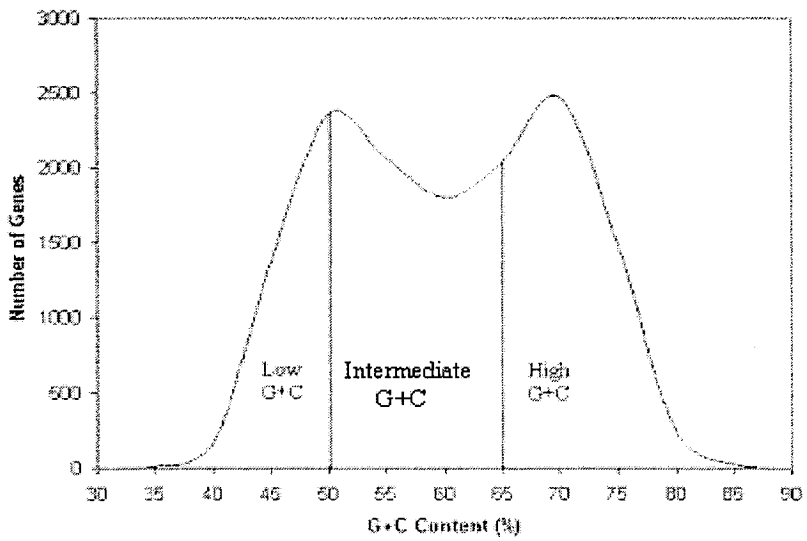


Fig. 5.1 The G+C content of 14005 rice genes shows a two modal distribution and was separated into low, intermediate and high GC classes.

Figure 5.2 shows GC disparity plots for high GC set, intermediate GC set and low GC set, respectively. GC disparity = 0 means the GC content at that codon position is same as the average GC content of the gene. Figure 5.2A indicates the distribution of GC1 is nearly centered at point 0 on the X axis, while the distributions of GC2 and GC3 are shifted leftward and rightward, respectively, which means GC2 is least G+C rich and GC3 is most abundant in the high G+C rice genes. Figure 5.2B also indicates GC3>GC1>GC2 and GC1 is almost centered at point 0 but the pick of GC3 is left shifted causing a lot of overlaps between GC1 and GC3. Figure 5.2C indicates GC3 is further left shifted and centered at point 0 on the X axis but still remains to the right of GC2 while GC1 is now shifted to the right of GC3. This indicates GC2 is also least rich but GC1 is most abundant in G+C for the low G+C rice genes.

5.4.2 Relative synonymous codon usage

The RSCU is a normalized synonymous codon frequency so that it is not affected by different amino acid compositions. Table 5.1A and B lists the cumulative RSCU values for all 14005 rice genes and 25625 *Arabidopsis* genes, respectively. We see that in the rice data (5574687 codons in total) 29 out of 30 third synonymous sites cytidine is more abundant than uridine and guanosine is more abundant than adenosine. The only exception is CCC which is less frequent than CCU for coding proline. A similar trend was observed (Wang, Shi & Hao, 2002) with a much smaller rice codon usage data (27910 codons) from the Codon Usage Database (http://www.kazusa.or.jp/codon/). In the *Arabidopsis* data, however, this relationship (C>U and G>A) only exists in 4 out of 30 pairs of the synonymous third positions. These results indicate that the rice genome prefers to use C and G and *the Arabidopsis* genome prefers to use A and U at the third codon position.

The average RSCU values calculated for all rice genes ignored the large GC difference between different sets of rice genes. RSCU values were then calculated for the three GC sets of rice genes (Table 5.1C1, C2 and C3). All 30 pairs of the synonymous third codon sites in both high GC set and intermediate GC set show preference of C to U and G to A. The preference is particularly strong in the high GC set and it is moderate in the intermediate GC set. In contrast only 5 pairs of the synonymous third sites show moderate C and G rich preference in the low G+C set.

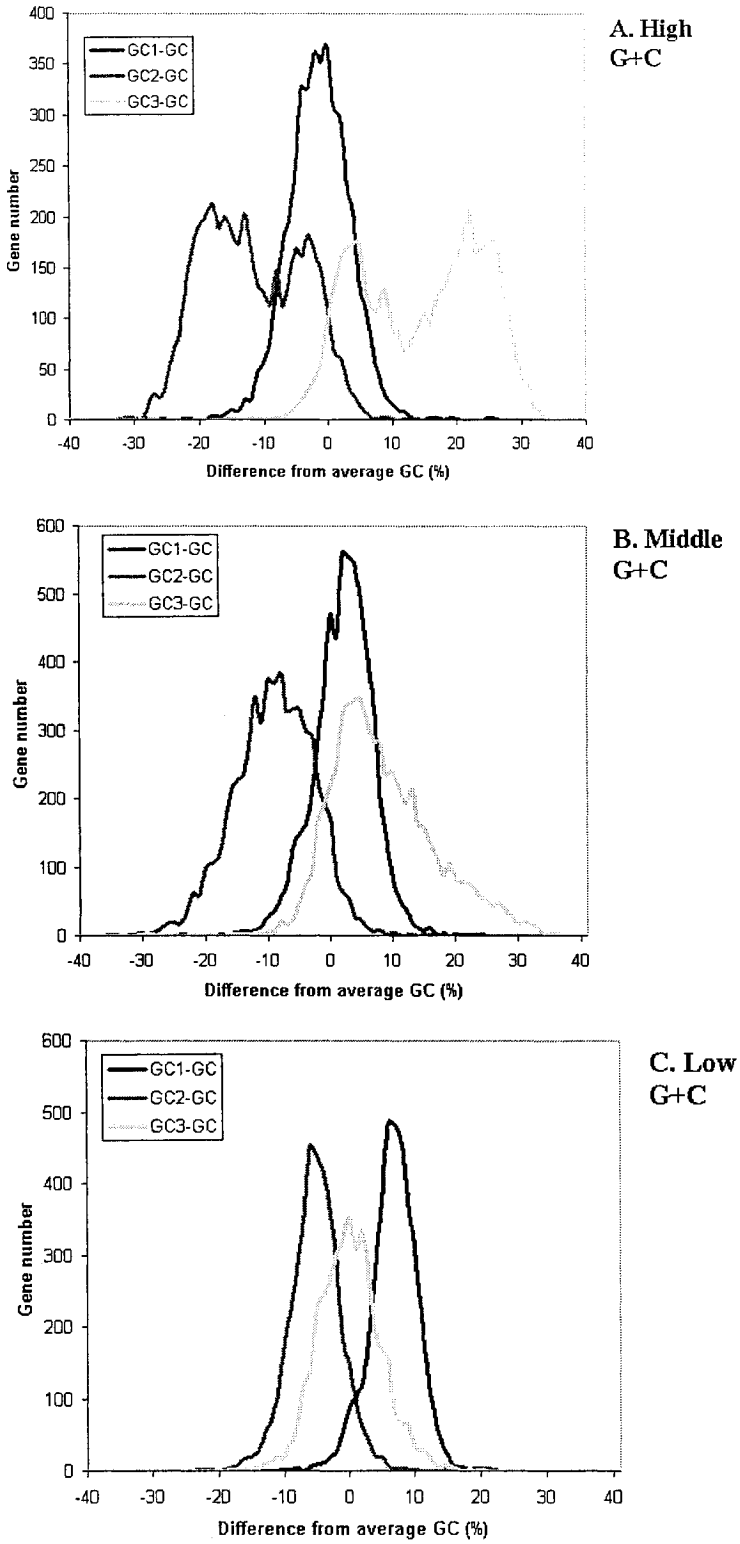


Fig. 5.2 The GC content disparity plot for rice genes. A) GC3 > GC1 > GC2 in high GC genes. B) The pattern GC3 > GC1 > GC2 is also seen in intermediate GC genes. C) GC1 > GC3 > GC2 in low GC genes.

Table 5.1 A and B) Cumulative RSCU for 14005 rice and 25625 *Arabidopsis* genes. **C1, C2, C3**) RSCU for the three GC sets of rice genes. NUA and NCG codons (discussed in the text) are highlighted.

		A) Rice		B) *Arabidopsis*		C1) LowGC Rice		C2) Intermediate GC Rice		C3) High GC Rice	
Codon	AA	ObsFreq	RSCU	ObsFreq	RSCU	ObsFreq	RSCU	ObsFreq	RSCU	ObsFreq	RSCU
UGA	*	6052	1.344	11033	1.292	1994	1.336	2521	1.316	2011	1.459
UAG	*	4138	0.919	5396	0.632	1223	0.819	1781	0.93	1365	0.99
UAA	*	3314	0.736	9196	1.077	1262	0.845	1445	0.754	760	0.551
GCA	A	97331	0.744	193892	1.11	60589	1.123	39003	0.736	11553	0.277
GCU	A	109246	0.835	301257	1.724	67046	1.243	43541	0.822	13032	0.313
GCG	A	146146	1.117	95739	0.548	35681	0.662	56909	1.074	72756	1.746
GCC	A	170516	1.304	108005	0.618	52415	0.972	72503	1.368	69332	1.664
UGC	C	69308	1.322	80598	0.796	28905	1.128	29762	1.385	19762	1.784
UGU	C	35549	0.678	121871	1.204	22366	0.872	13225	0.615	2391	0.216
GAU	D	139188	0.94	416782	1.375	92156	1.2	50370	0.864	10360	0.338
GAC	D	157058	1.06	189507	0.625	61442	0.8	66270	1.136	50952	1.662
GAG	E	215032	1.28	358969	0.954	98499	1.114	88655	1.353	58855	1.712
GAA	E	121067	0.72	393962	1.046	78341	0.886	42425	0.647	9887	0.288
UUU	F	74411	0.751	250309	1.056	52414	0.95	25023	0.633	3296	0.189
UUC	F	123791	1.249	223706	0.944	57951	1.05	54022	1.367	31543	1.811
GGA	G	89856	0.832	259138	1.465	50303	1.064	35377	0.823	15047	0.468
GGU	G	82299	0.762	237467	1.342	50508	1.068	32708	0.761	10744	0.334
GGG	G	95488	0.884	112713	0.637	38806	0.821	38073	0.886	32274	1.004
GGC	G	164297	1.521	98274	0.556	49547	1.048	65776	1.53	70552	2.194
CAC	H	78737	1.104	94725	0.751	31046	0.866	34954	1.194	23651	1.644
CAU	H	63845	0.896	157465	1.249	40632	1.134	23602	0.806	5123	0.356
AUC	I	110392	1.385	200698	1.021	51152	1.079	50384	1.609	25077	2.429
AUU	I	80294	1.008	242997	1.236	57104	1.205	27360	0.874	3582	0.347
AUA	I	48397	0.607	146215	0.744	33970	0.717	16214	0.518	2312	0.224
AAG	K	178251	1.326	360788	1.016	99204	1.235	73224	1.42	31057	1.747
AAA	K	90682	0.674	349573	0.984	61412	0.765	29891	0.58	4506	0.253
UUA	L	34690	0.421	147322	0.841	24617	0.559	9992	0.303	1135	0.07
UUG	L	81823	0.994	237822	1.357	53699	1.22	30250	0.919	6979	0.427
CUA	L	43089	0.523	113640	0.649	27629	0.628	16097	0.489	3459	0.212
CUC	L	137895	1.674	173537	0.99	50807	1.154	61534	1.869	48195	2.952
CUG	L	113446	1.378	111451	0.636	50298	1.143	48625	1.477	32085	1.965
CUU	L	83159	1.01	267620	1.527	57021	1.296	31073	0.944	6111	0.374
AUG	M	129867	1	271189	1	68471	1	52434	1	25411	1
AAU	N	83008	0.885	258847	1.06	58442	1.055	26501	0.744	3577	0.282
AAC	N	104533	1.115	229749	0.94	52330	0.945	44695	1.256	21835	1.718
CCG	P	99607	1.236	91534	0.689	26278	0.745	42439	1.234	43240	2.063
CCU	P	75058	0.931	203205	1.529	44203	1.254	30860	0.897	9141	0.436
CCA	P	78794	0.978	179280	1.349	45851	1.3	32134	0.934	9623	0.459
CCC	P	68900	0.855	57600	0.433	24697	0.7	32185	0.935	21855	1.042
CAG	Q	116456	1.203	165222	0.859	58084	1.091	49436	1.275	26601	1.642
CAA	Q	77166	0.797	219549	1.141	48403	0.909	28102	0.725	5806	0.358
CGG	R	78739	1.176	54009	0.542	22240	0.807	31382	1.139	33686	1.72
CGC	R	92544	1.382	41164	0.413	26674	0.967	39726	1.442	38994	1.991
CGA	R	41111	0.614	71172	0.715	17002	0.617	18672	0.678	9104	0.465
AGA	R	59100	0.883	213059	2.14	36048	1.307	20636	0.749	6546	0.334
AGG	R	86730	1.295	121522	1.22	41472	1.504	35487	1.288	21513	1.098
CGU	R	43579	0.651	96481	0.969	22002	0.798	19407	0.704	7676	0.392
AGU	S	48147	0.677	161503	0.973	32140	0.894	16453	0.585	3032	0.203
AGC	S	88186	1.24	125046	0.754	37388	1.04	36889	1.313	25128	1.682
UCU	S	68302	0.961	280101	1.688	44521	1.238	23790	0.847	5499	0.368
UCG	S	67361	0.948	101462	0.612	22066	0.614	28691	1.021	24824	1.662
UCA	S	65321	0.919	206240	1.243	43762	1.217	23468	0.835	4456	0.298
UCC	S	89233	1.255	121086	0.73	35924	0.999	39327	1.399	26698	1.787
ACA	T	64282	0.934	177876	1.256	42277	1.25	24059	0.842	4344	0.301
ACC	T	85386	1.24	111114	0.785	34347	1.015	39554	1.384	23763	1.647
ACG	T	65370	0.95	84494	0.597	18793	0.555	27842	0.974	26032	1.804
ACU	T	60334	0.876	192935	1.362	39923	1.18	22860	0.8	3582	0.248
GUG	V	133996	1.452	191093	1.027	57710	1.24	55839	1.481	40215	1.937
GUC	V	110668	1.199	137771	0.74	42777	0.919	49710	1.318	34837	1.678
GUA	V	38574	0.418	114363	0.615	25732	0.553	13897	0.369	2656	0.128
GUU	V	85927	0.931	301126	1.618	59970	1.288	31398	0.833	5323	0.256
UGG	W	78353	1	139910	1	37498	1	31502	1	18387	1
UAU	Y	56361	0.798	168645	1.063	40768	1.008	19056	0.675	1929	0.169
UAC	Y	84907	1.202	148515	0.937	40149	0.992	37371	1.325	20864	1.831

The difference in the base usage in the synonymous third codon sites among the high GC, intermediate GC and low GC rice genes and the *Arabidopsis* genes parallels the direction of average GC content of the three sets of rice genes and the *Arabidopsis* genes. We were concerned that the difference between the codon usage patterns of rice and *Arabidopsis* gene might simply reflect a difference in gene content between the two genomes. To investigate this possibility we first did a BLAST search to identify 7160 pairs of homologous rice and *Arabidopsis* genes (see section 5.3.2). The rice homologous genes were further separated as high GC (G+C > 65%), intermediate GC (between 50 and 65%) and low GC (G+C ≤ 50%) and the *Arabidopsis* homologs for each subset were identified. The three sets of rice homologous genes have an average GC content close to the average GC content of all high GC, intermediate GC and low GC genes, respectively. The three sets of *Arabidopsis* homologs have similar and lower average GC contents, but the set homologous to high GC rice genes has slightly higher mean GC than the set homologous to intermediate GC rice genes, and the later has slightly higher mean GC than the set homologous to low GC rice genes (Table 5.2). Figure 5.3 plots GC content of the three sets of rice genes against *Arabidopsis* homologs. The GC content of the rice genes varies from 29% to 80%, while the corresponding range for the *Arabidopsis* homologs is only 35% to 59%. This indicates the rice genes are more variable in GC content than their *Arabidopsis* homologs.

Table 5.2 Gene number and average GC content (%) of high, intermediate and low GC sets of rice genes and *Arabidopsis* homologs.

	riceHighGC	*Arab.* homolog	**rice Interm.GC**	*Arab.* homolog	**rice LowGC**	*Arab.* homolog
Gene number	1555	1555	2939	2939	2666	2666
Mean GC	69.0	46.9	56.8	45.6	45.7	43.9
S.D.	2.54	3.28	4.54	2.92	2.80	2.72

We then computed RSCU of the three sets of rice genes, along with their *Arabidopsis* homologs (data not shown). All 30 pairs of synonymous third codon sites in the high GC rice set show strong preference of C and G to U and A, while only 8 pairs of synonymous third sites in the *Arabidopsis* homologs show a weaker C and G preference. The intermediate GC rice set also shows moderate preference of C and G to U and A in

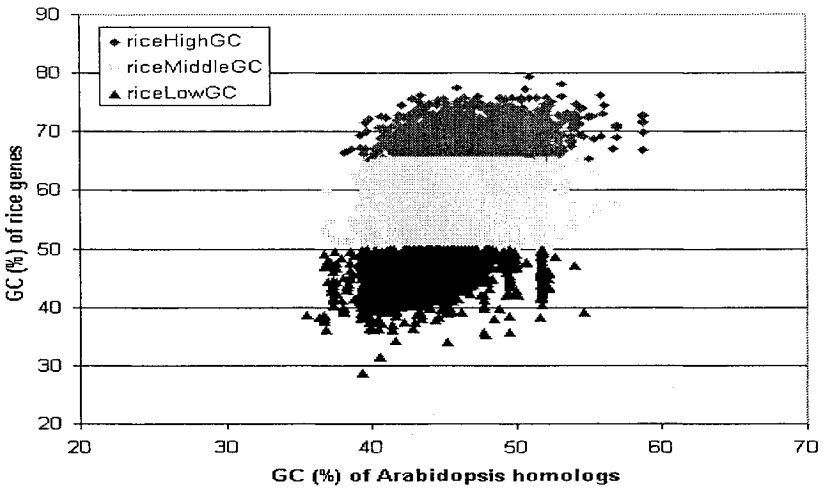


Fig. 5.3 GC content of high, intermediate and low GC sets of rice genes plotted against *Arabidopsis* homologs.

all 30 pairs of synonymous third codon sites while its *Arabidopsis* homologs have this preference only in 7 synonymous pairs. The low GC rice set only shows the C and G preference in 5 synonymous pairs and its *Arabidopsis* homologs have 4 synonymous pairs in this preference. These demonstrate that the difference in the base usage preference between rice and *Arabidopsis* genes is not due to the difference in their gene contents but due to the difference in their GC contents.

RSCU for all rice genes and *Arabidopsis* genes and RSCU for the three sets of the rice genes (Table 5.1) and for their *Arabidopsis* homologs also show that for both rice genes and *Arabidopsis* genes NUA codons are least frequent. The scarcity of NUA codons is most obvious in the rice high GC genes (Table 5.1C3). The NCG codon, however, is not as infrequently used in both rice and *Arabidopsis* gene sets. The very low frequency of NUA codons and moderate frequency of NCG codons in high GC rice genes reflect a strong bias for codon ending G or C in these genes. In summary, the RSCU values clearly indicate the base usage of synonymous third codon positions of rice genes is determined by the G+C content of the genes.

5.4.3 Codon usage entropy

Codon usage entropy was calculated for each gene in the three sets of rice genes (high, intermediate and low GC classes). The average entropy (mean ± S.E.) for the high GC genes (1.13 ± 0.005 bits) was much lower than that for the intermediate GC genes (1.47 ±

0.002 bits) and low GC genes (1.47 ± 0.002 bits), indicating the codon usage of high GC genes are more biased than the rice genes of lower GC. The range of the entropy variation in the high and intermediate GC classes (1.34 and 1.38 bits, respectively) is higher than the range seen in the low GC class (0.85 bit), which suggests that the later class of genes is more homogenous in codon bias.

The codon entropy was then calculated for the three sets of rice genes and their *Arabidopsis* homologs, respectively (Table 5.3). The results show that rice high GC genes have much lower codon usage entropy not only compared to their *Arabidopsis* homologs but also to the other rice genes (of intermediate or low GC). The rice intermediate and low GC genes have average entropy fairly close to their *Arabidopsis* homologs, respectively. The three sets of *Arabidopsis* gene homologs have similar average codon entropy and small range of variations, indicating *Arabidopsis* codon usage is rather homogenous. In contrast, the ranges of entropy variations are much wider in the intermediate and high GC rice genes, indicating they are heterogeneous in codon usage.

Table 5.3 Average and range of codon usage entropy (in bits) of high, intermediate and low GC sets of rice genes and *Arabidopsis* homologs.

	riceHighGC	*Arab.* homolog	**rice Interm.GC**	*Arab.* homolog	**rice LowGC**	*Arab.* homolog
Gene number*	1555	1555	2939	2939	2666	2666
Actual gene number**	1390	1422	2627	2695	2523	2581
Codon entropy	0.99	1.48	1.42	1.47	1.50	1.49
Entropy range***	1.16	0.78	1.36	0.79	0.85	0.79

* original number of genes in each class.
** The number of genes that codon entropy can be calculated.
** Maximum entropy – minimum entropy in each class.

5.4.4 Effective number of codons

The effective number of codons (Nc) is a commonly used measurement to quantify codon usage bias of a gene. As suggested by Wright (1990) the Nc values are viewed on an Nc plot, which plots Nc against GC3s of the genes. Figure 5.4 is an Nc plot for all 14005 rice genes with 88 ribosomal protein genes highlighted. The GC3s vary widely from 18% to

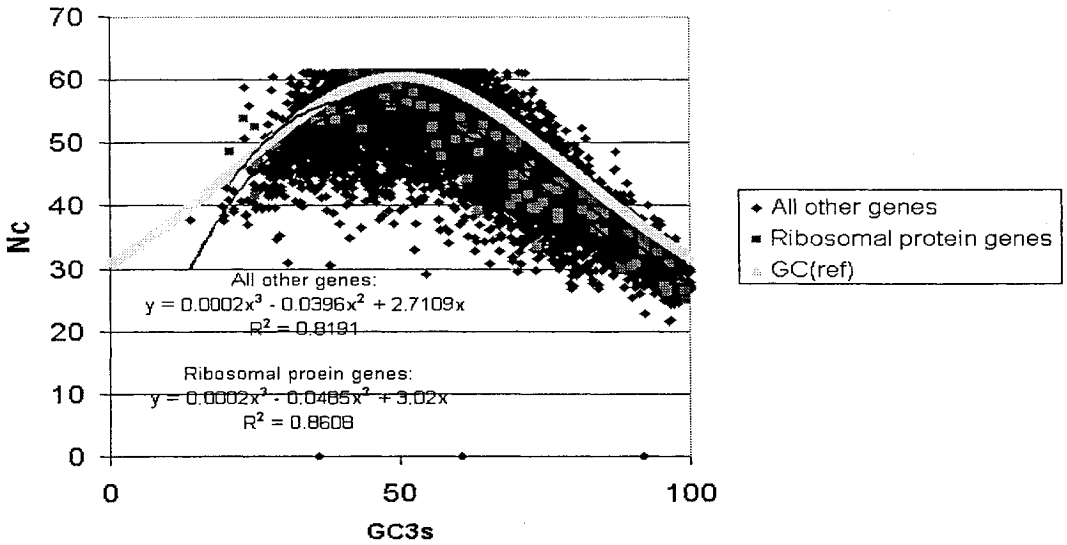


Fig. 5.4 Nc plot for 14005 rice genes with ribosomal protein genes highlighted. GC(ref) is the expected position of genes whose codon usage is only determined by GC3s. The two black curves are the best-fit lines for ribosomal genes and all other genes, respectively.

100%. The plot contains a reference line (GCref) showing the expected position of genes whose codon usage is only determined by variation in GC3s, which is an approximate upper limit for the value of Nc (Wright, 1990). The majority of the genes, including those encoding ribosomal proteins, are densely located along or not far from the reference line. The ribosomal protein genes are distributed according to their GC3s and they do not form a cluster. The two polynomial lines fit the data very well for both ribosomal protein genes ($R^2 = 0.86$) and all other genes ($R^2 = 0.82$), indicating GC3s is a dominant factor in rice codon bias.

Figure 5.5 is an Nc plot for all 25625 *Arabidopsis* genes with 255 ribosomal protein genes highlighted. The GC3s have a range from 23% to 70%, with the majority in the range from 35% to 45%. Some of the genes, including many of the ribosomal protein genes, lie well below the reference line (GCref) suggesting that GC3s only explains some of the variation in Arabidopsis codon usage. Indeed a polynomial line best fitting the data only achieves a $R^2 = 0.14$. The very biased codon usage relative to the GC3s in the ribosomal protein genes suggests gene expression is important in shaping *Arabidopsis* codon bias. This agrees with a previous study that clustered 718 *Arabidopsis* genes into two codon usage groups, one of which contains genes of high expression including 17 of

18 ribosomal protein genes in the dataset and another group contains low expression genes such as regulatory genes (Mathe et al., 1999).

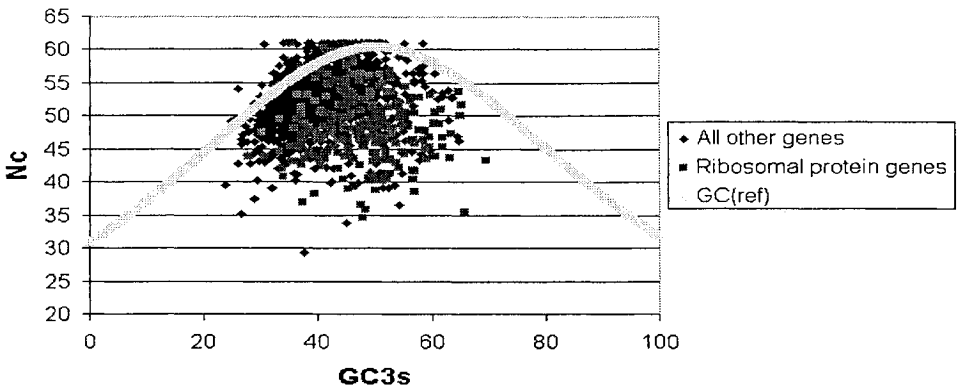


Fig. 5.5 Nc plot for 25625 Arabidopsis genes. The ribosomal protein genes (in red squares) and some other genes are much lower than the GC3 reference line, indicating high codon bias.

To get rid of the effect of gene content difference on the Nc plot, an Nc plot for 7160 rice genes homologous to *Arabidopsis* genes and a plot of that for the corresponding 7160 *Arabidopsis* homologs were made, respectively (Figs. 5.6 and 5.7). As for the original 14005 rice genes, the rice homologous genes are more stretched along the X axis and the genes are not far from the reference line. The *Arabidopsis* homologous genes, like the total *Arabidopsis* genes, are more focused and some genes are far from the reference line. A polynomial curve fits very well in the rice homologs ($R^2 = 0.91$) but it fits poorly in the Arabidopsis homologs ($R^2 = 0.13$). These again suggest GC3s is a dominant factor in shaping rice codon bias but not in the *Arabidopsis* genes.

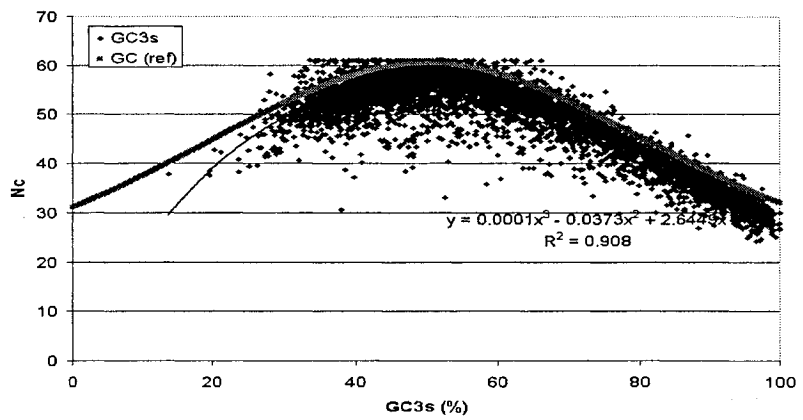


Fig. 5.6 Nc plot for 7160 rice genes homologous to Arabidopsis genes. The polynomial line accounts for 91% of the variation in the data.

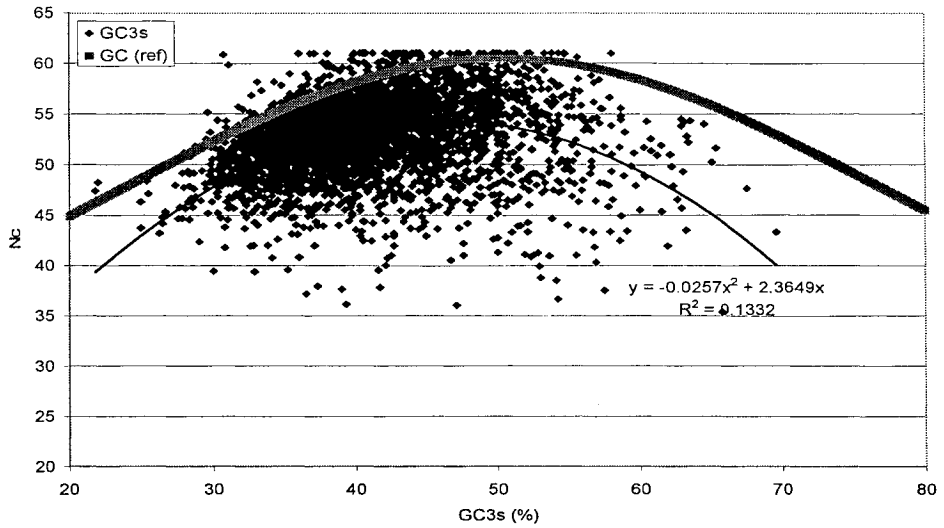


Fig. 5.7 Nc plot for 7160 Arabidopsis genes homologous to rice genes. The polynomial line only accounts for 13.3% variation in the data.

5.4.5 Correspondence analysis

While an Nc plot can indicate the heterogeneity of codon usage among genes it cannot show the underlying distribution of codons. Correspondence analysis has the advantage of showing both gene and codon distributions on a two dimensional space, the first two axes of which capture the largest variation in the data (Greenacre, 1984). Figure 5.8 shows a correspondence analysis of the value of relative synonymous codon usage in 14005 rice genes. The origin in Fig. 5.8A represents the average RSCU for all genes, with respect to the first two axes. The distance between genes on this plot is a reflection of their dissimilarity in RSCU, with respect to the two axes. The two axes account for 36.9% and 6.9% of the variations in the data, respectively. The genes are separated into high, intermediate and low GC classes along the primary axis. There are some overlaps between the high and intermediate GC gene classes and some overlaps between the intermediate and low GC classes but no overlap between the high and low GC classes along the primary axis (Fig. 5.8A). The underlying codon distribution in the correspondence analysis also indicates a clear separation of C/G-ending codons and A/U-ending codons along the primary axis (Fig. 5.8B).

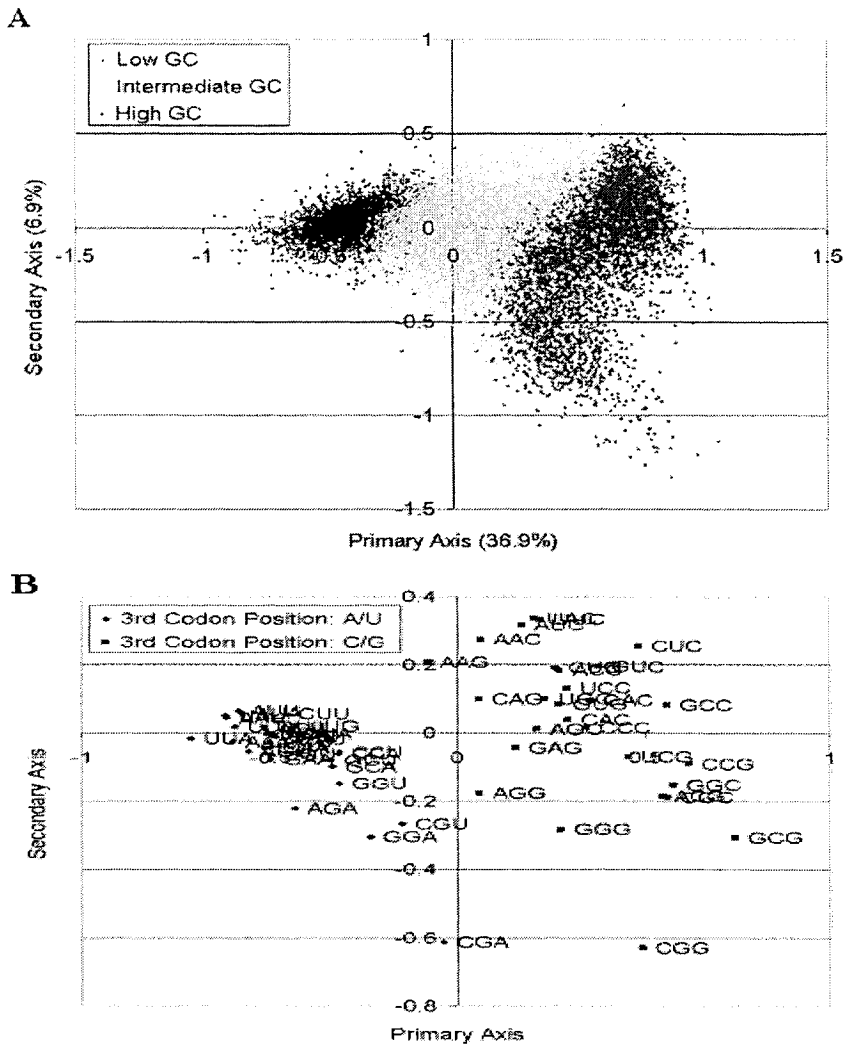


Fig. 5.8 Correspondence analysis of RSCU of 14005 rice genes. A) the distribution of the genes. Each point is a gene, colored as gene classes: high GC genes (pink), intermediate GC genes (yellow) and low GC genes (blue). B) the distribution of the codons. C/G-ending codons and A/U-ending codons are marked as red square and blue diamond, respectively.

Furthermore, a plot of GC1, GC2 and GC3 against gene location on the primary axis shows they all are correlated with the primary position ($p < 0.00001$ in all cases) and GC3 has the strongest correlation (Fig. 5.9). Positions on the secondary axis show a moderate correlation with both GC2 ($R = 0.69$, $p < 0.00001$) and GC1 ($R = 0.49$, $p < 0.00001$) but no correlation with GC3 ($R = 0.08$). The correlations of GC3 and also GC1 and GC2 with gene positions on the first two axes demonstrate a generalized nucleotide bias affecting codon usage in the rice genes.

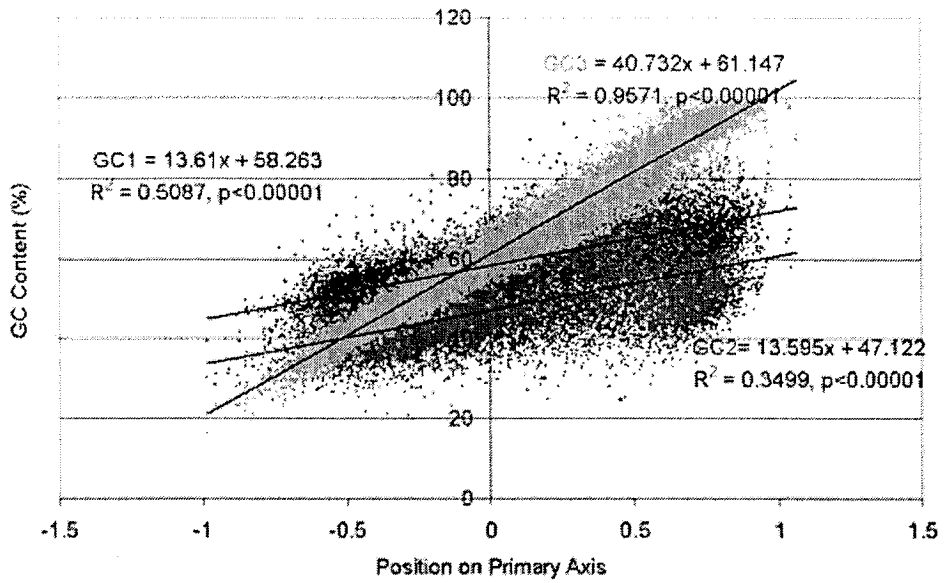


Fig. 5.9 Correlation between GC1, GC2 and GC3 with gene locations on the primary axis of the correspondence analysis in Fig. 5.8A.

Figure 5.10 shows a correspondence analysis of the values of RSCU in 7160 homologous gene pairs between rice and *Arabidopsis* (total 14320 genes). The primary axis accounts for 40.2% of the variations in the data whereas the next three axes only account for 4.3, 3.9 and 3.3%, respectively. The distribution of genes (Fig. 5.10A) is clearly determined by the distribution of G+C contents in the genes themselves as the high GC rice genes and low GC rice genes/*Arabidopsis* genes are located at the two ends of the primary axis with intermediate GC rice genes in the middle. There are a lot of overlaps between the low GC rice genes and *Arabidopsis* genes, suggesting codon usage is greatly influenced by G+C content rather than the species. The distribution of codons (Fig. 5.10B) shows a clear separation of C/G-ending codons and A/U-ending codons along the first axis in the direction of GC-rich rice genes and AT-rich *Arabidopsis* genes. A close look at the distribution of codons revealed that the G+C content of the codons follows very closely the G+C content of the genes. For instance, of the 8 codons with 100% G+C content five of them (GCC, GGC, CCG, CGC, GCG) are exclusively located in the high GC rice gene cluster, two of them (CCC and CGG) located in the junction of high GC and intermediate GC rice gene clusters and one (GGG) in the intermediate GC rice gene cluster. Codons that have 2 Gs and Cs out of the three codon positions are located in the intermediate GC

gene cluster. Codons have 2 or 3 As and Us are located in the low GC rice gene and *Arabidopsis* gene cluster.

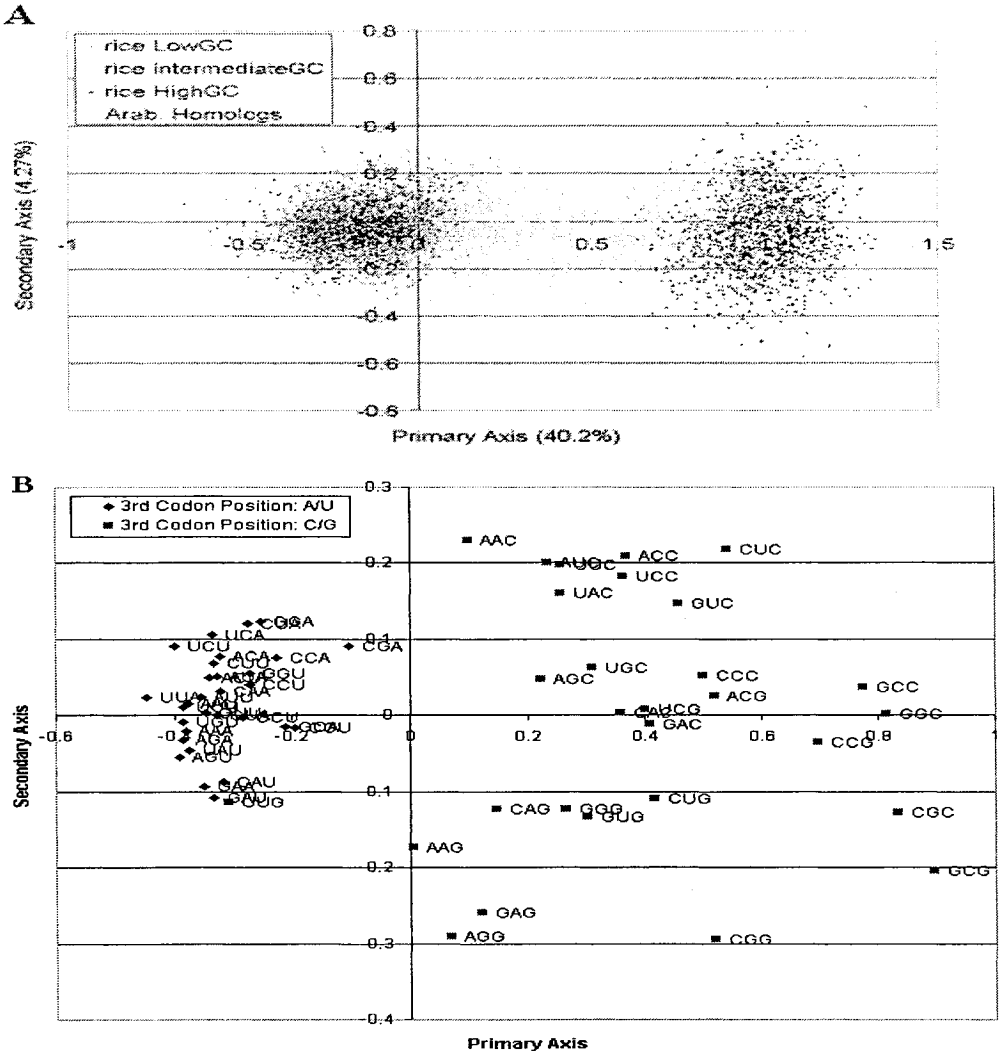


Fig. 5.10 Correspondence analysis of relative synonymous codon usage of 7160 rice genes and 7160 *Arabidopsis* homologs. A) gene distribution. Each point is a gene, colored as gene classes: high GC rice genes (pink), intermediate GC rice genes (yellow), low GC rice genes (blue) and *Arabidopsis* homologs (green). B) codon distribution. C/G-ending codons and A/U-ending codons are marked as red square and blue diamond, respectively.

The correlation between the gene distribution and G+C content is further reinforced when we plot the G+C content of individual genes against their position on the primary axis of the correspondence analysis (Fig. 5.11). The rice genes present a very

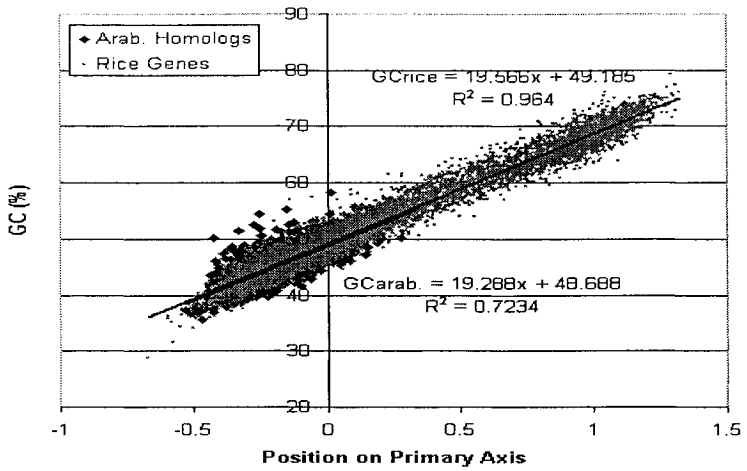


Fig. 5.11 Correlation of G+C content of 7160 rice genes (red points) and 7160 *Arabidopsis* homologs (blue diamonds) with their positions on the primary axis of the correspondence analysis (Fig. 5.10), respectively.

high correlation between GC content and gene position on the primary axis (regression R^2 = 0.96, P < 0.00001), while their *Arabidopsis* homologs present a relatively weaker correlation (regression R^2 = 0.72, P < 0.00001). Furthermore, GC1 and GC2 also show moderate correlation with gene position on the primary axis (P < 0.00001 in both cases; Fig. 5.12). Taken together, the correspondence analyses indicate that G+C content is the primary determinant of codon usage among the genes. The changes in nucleotide content result in very significant changes in the codon usage patterns of the rice genes, especially in the high G+C rice genes.

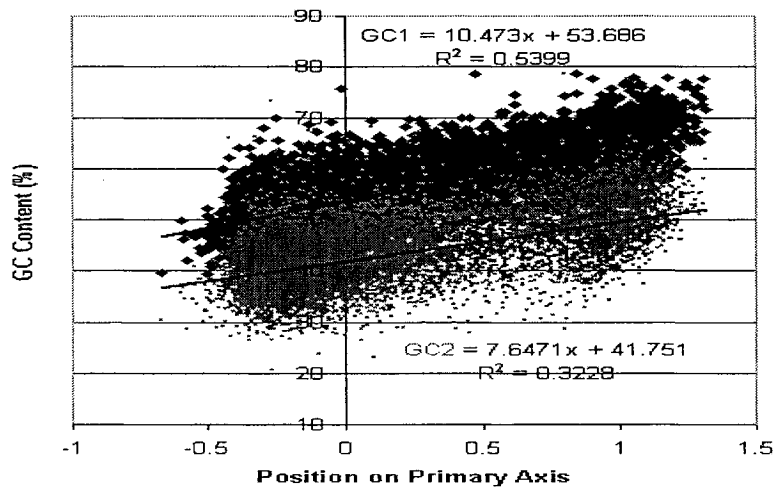


Fig. 5.12 Correlation of GC1 and GC2 of 14320 rice and *Arabidopsis* homologous genes with their position on the primary axis of the correspondence analysis (Fig. 5.10).

5.4.6 Codon adaptation index

The analyses described above demonstrate that GC content is a dominant variable for rice codon usage and that translational bias plays a minor, if any, role; this is suggested by the fact that the highly expressed ribosomal protein genes do not show high codon bias (Fig. 5.4). The later point is further reinforced when we analyzed a revised codon adaptation index (Carbone, Zinovyev & Kepes, 2003) of the rice genes. Unlike the traditional CAI, this revised method does not need a predefined reference set that consists of highly biased genes which are known to have high expression levels such as ribosomal protein genes and other protein translation-related genes. Instead, it automatically determines 1% of genes with highest CAI values and uses them as a reference set to calculate CAI of other genes. If the majority of the genes in the reference set are ribosomal protein genes and other known highly expressed genes, then the codon usage is driven by selection for gene expression is suggested (Carbone, Zinovyev & Kepes, 2003). Applying the new method to the 14005 rice gene data, after 20 iterations the program converged and the CAI values fixed at stable values. A reference set of 140 genes was extracted. The mean and standard deviations of GC and GC3 of the reference set are 0.69 ± 0.03 and 0.98 ± 0.01, respectively, indicating they actually consist of high GC genes. Of the 140 genes only 6 are ribosomal protein genes out of 88 ribosomal protein genes in the rice data. Other genes in the reference set include heat shock proteins, pathogenesis-related protein, nitrate transporter, ribonuclease and catalase, etc. A plot of GC3 vs. the CAI values for all rice genes indicates a strong linear correlation (the R^2 value is almost 1) and the ribosomal protein genes are also linearly distributed on the line along the whole range of the distribution (Fig. 5.13). This indicates GC3, rather than gene expression, dominates codon bias of rice genes. Moreover, GC1 and GC2 also have weak correlation with CAI ($p < 0.00001$ in both cases; Fig. 5.14), indicating a generalized nucleotide bias affecting the codon usage.

The CAIjava program was also applied to 6627 genes on *Arabidopsis* chromosome 1 and converged after 7 iterations. The reference set contains 66 genes that have an average GC of 39.7% and average GC3 of 34.3%. The grand mean GC and GC3 of all of the 6627 genes are 44.5% and 42.8%, respectively. This indicates the CAIjava method found a biased reference set of low GC genes. The data set contains 61 ribosomal

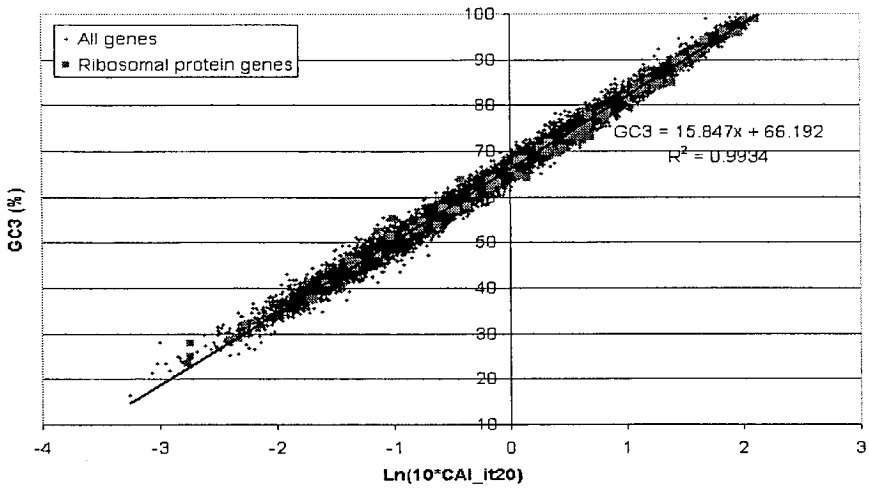


Fig. 5.13 GC3 of 14005 rice genes including ribosomal protein genes are linearly correlated with CAI. The CAI values, calculated from the 20th iteration when the CAIjava program converged, were logarithm-transformed so that they are closer to a normal distribution.

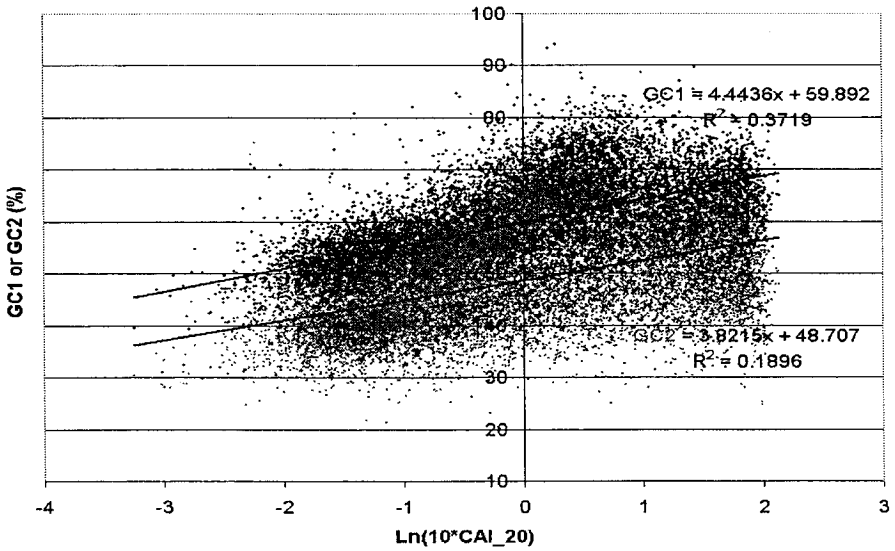


Fig. 5.14 GC1 and GC2 of 14005 rice genes show weak positive correlation with CAI (in logarithm-transformed form).

genes that have mean GC and GC3 of 47.1% and 49.3%, respectively. They are higher than average GC and GC3 of all genes and so their average CAI value (0.52) is a little lower than the mean CAI of all genes (0.53) as the reference set detected low GC genes as biased codon usage. A scatter plot of GC3 vs. CAI for the 6627 *Arabidopsis* chromosome 1 genes is shown in Figure 5.15. This regression R^2 is only 0.53, indicating

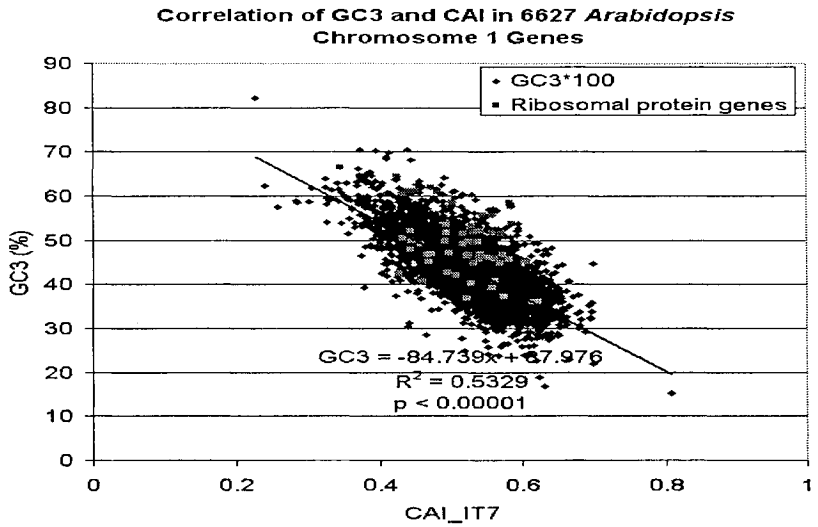


Fig. 5.15 Correlation of GC3 and CAI (at iteration 7) in 6627 *Arabidopsis* chromosome 1 genes. The ribosomal protein genes are highlighted.

CAI is only half explained by the GC3 in the *Arabidopsis* genes. Furthermore, both rice and *Arabidopsis* homologous genes are subject to CAIjava analyses, respectively. The CAIs of rice homologous genes present similar perfect correlation with GC3 as seen in all rice genes and the *Arabidopsis* homologs present a CAI-GC3 correlation similar to all *Arabidopsis* genes. These demonstrate rice codon usage is almost entirely dominated by GC3 while other factors, in addition to GC3, shape *Arabidopsis* codon usage.

5.5 Discussion

We have examined codon usage in 14005 rice genes and 25625 *Arabidopsis* genes and 7160 pairs of homologous genes between the two genomes. The rice data was further divided into high, intermediate and low GC sets. The pattern of relative use of synonymous codons was shown to differ among the three sets of genes and between the rice and *Arabidopsis* homologous genes, primarily in the use of G+C in the synonymous third codon position, rather than between the species. The high GC rice genes have high frequency of G and C at the synonymous third sites (GC3s) and the *Arabidopsis* genes and low GC rice genes have high frequencies of A and T at the GC3s, while the intermediate GC rice genes have a GC3s frequency between the two high GC and high AT groups, consistent with their intermediate GC content.

The codon usage entropy analyses indicate: (1) the rice genes have a wide range of codon usage, while the codon usage of the *Arabidopsis* genes are rather homogenous. (2) the high GC rice genes are very biased in codon usage, while the low GC rice genes have similar low codon bias as the *Arabidopsis* genes. The distribution of rice and *Arabidopsis* genes on the primary axis in the correspondence analyses of the RSCU values is strongly correlated with GC and GC3 contents of the genes (Figs. 5.9; 5.11). Furthermore, the gene distributions on the primary and secondary axes of the correspondence analyses are correlated with GC1 and GC2 (Figs. 5.9; 5.12). These provide strong evidence for a generalized nucleotide bias affecting the codon usage. Previously it was shown that variation in GC3s is also a major source of variation in the codon usage of *Zea mays* (Fennoy & Bailey-Serres, 1993). The current comparative codon usage study between rice and *Arabidopsis* homologous genes indicates the main difference in the codon usage between monocotyledons and dicotyledons is the average GC3s of the genes.

The variation in GC3s of rice genes could reflect regional heterogeneity in GC content in the rice genome, or selection constraints within genes operating at any level of gene expression. Unlike *Arabidopsis* genes that a high codon bias is present in ribosomal protein genes (Fig. 5.5; Mathe et al., 1999), the ribosomal protein genes are not all in high codon bias in the Nc plots (Fig. 5.4). As the ribosomal protein genes are presumably highly expressed, this suggests that selection constraint, if any, plays a minor role in shaping rice codon usage. However, one limitation of Nc is that it measures the departure from uniform codon usage and this later assumption is not always desirable (Novembre, 2002). Previous investigations have suggested the presence of randomly dispersed 50 to 100 kB GC-rich isochores among more AT-rich chromosomal regions in plants, particularly in monocots (Matassi et al., 1989). We have shown that the rice genome has a large heterogeneity in GC content (Fig. 5.1). These suggest a null distribution of codon usage is nonuniform, which violates the assumption for calculating the Nc. In such cases background nucleotide composition should be taken into account when calculating codon bias (Urrutia & Hurst, 2001; Novembre, 2002). Here, we used Ncprime, a modified Nc, that account for background nucleotide composition to calculate the codon bias (Novembre, 2002). For the 14005 rice genes, we extracted 9620 genes that have intron

sequences that are longer than 500 bases and the nucleotide composition of the introns were used as background nucleotide composition for calculating Ncprime for each of the 9620 genes. The average Ncprime for the 9620 genes is 46.2 ± 0.1, down from the original Nc of 49.6 ± 0.08. There are 39 ribosomal protein genes in this data set. The average Ncprime of the ribosomal protein genes is 43.1 ± 1.58, down from the original Nc of 46.6 ± 1.21. This result indicates average codon bias of ribosomal proteins is not significantly higher than average codon bias of all genes, even the background nucleotide composition has been taken into account. This is further supported by the analysis of CAI using a revised CAI method that extracted a reference set of codon bias from the rice genes and the reference set contained few ribosomal protein genes.

To further investigate whether variation in GC3s may be due to regional GC content effect, we compared GC3s to G+C content of introns and GC3s to GC1 and GC2 of the coding regions for the 9620 rice genes (Fig. 5.16). Both introns and the coding regions (GC1 and GC2) have positive correlation with GC3s, but the intron GC and GC3 correlation is weaker (Pearson correlation $R = 0.31$, $p < 0.00001$) than GC1 and GC3 correlation ($R = 0.61$, $p < 0.00001$) or GC2 and GC3 correlation ($R = 0.46$, $p < 0.00001$).

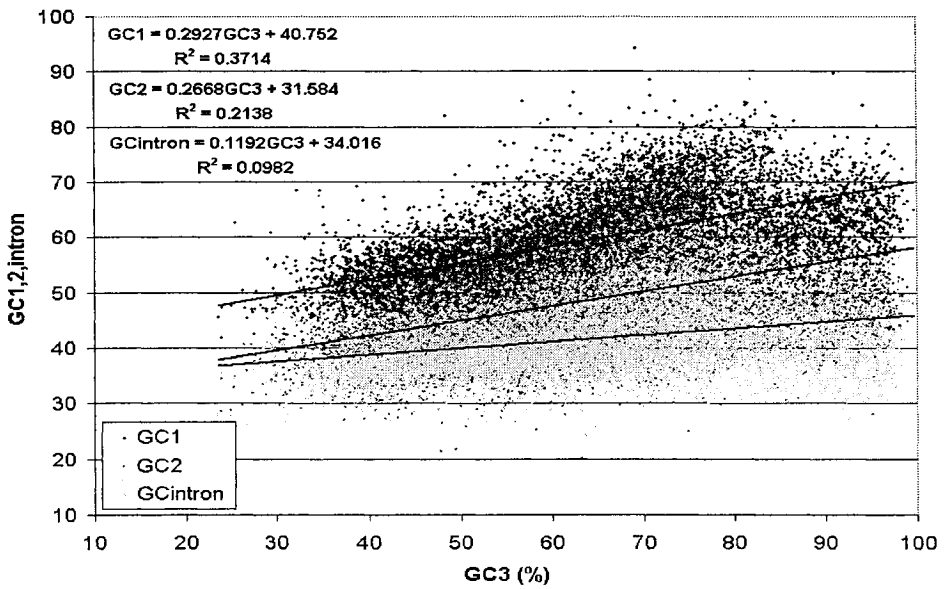


Fig. 5.16 Correlation of frequency of GC1 and GC2 in coding regions and GC frequency in introns, respectively, with GC3s for 9620 rice genes having an intron length greater than 500 bases. The least squares lines of best fit are plotted.

This result indicates regional compositional bias may not be solely responsible for the variation in GC3s. Obviously other factors may affect rice synonymous codon usage.

One of the factors to be considered is gene length, which has previously been found negatively correlated with codon usage in *C. elegans*, *Drosophila* and *Arabidopsis* (Duret & Mouchiroud, 1999). Correlation analysis of Nc, Ncprime and logarithm of protein length (number of codons) for the 9620 rice genes indicated a positive correlation between Nc and the length ($R = 0.15$, $p < 0.00001$) and between Ncprime and length ($R = 0.18$, $p < 0.00001$). This indicates in rice genome, as in *Arabidopsis*, *C. elegans* and *Drosophila*, gene length has a negative correlation with codon bias.

Another factor is selection constraint. If codon bias is due to selective pressures then we would expect genes with higher codon bias would have lower rates of synonymous substitutions (Urrutia & Hurst, 2001). Such an inverse correlation has been observed in bacteria (Sharp & Li, 1987b), Drosophila (Powell & Moriyama, 1997) and yeast (Urrutia & Hurst, 2001), the codon usage of which all subject to selection pressures. We already see above that selection for translation efficiency seems to play a minor, if any, role in rice codon usage. This is further reinforced by the fact that there is a positive correlation between codon bias and rate of synonymous substitutions (dS). We found a negative correlation between Nc and dS for 895 pairs of rice and *Arabidopsis* chromosome 4 genes ($R = -0.27$, $p < 0.00001$) and the less Nc the higher codon bias. And this effect is not due to the correlation of length and codon bias, as a multiple regression analysis of dS and gene length on Nc indicates Nc still has a negative correlation with dS ($p < 0.00001$).

In summary, we have found that the codon usage of rice genes is primarily determined by the GC content of the genes; not all variations in GC3s can be attributed to regional GC content; shorter genes have greater codon usage bias; the rate of synonymous evolution is higher in genes with greater codon bias, suggesting of nonexistence of selection on the codon usage, and this is supported by the fact that most of the ribosomal protein genes do not have higher codon bias as suggested in the Nc plots and CAI analyses.

Chapter 6
Conclusions

The primary objectives of this project were to identify the source of nucleotide bias and to understand the effects of the bias on protein sequence change and genome evolution. Natural selection, in the form of thermal adaptation, was thoroughly studied on 16S and 18S rRNA genes and, to a lesser extent, on the GC content of whole genomes and protein coding genes (ribosomal protein genes) in Chapters 2 and 3. The consequences of nucleotide bias were investigated by comparing amino acid usage, codon usage, amino acid substitution and nucleotide substitution patterns of homologous genes between rice and *Arabidopsis* genomes in Chapters 4 and 5. The results led to the following conclusions.

6.1 Thermal adaptation of rRNA genes and the genomes

The GC contents of structural RNAs, such as rRNA, show a positive correlation with growth temperature optima in prokaryotes, and this correlation is concentrated almost entirely within the double stranded stem regions of the rRNAs. Ribosomal RNA stem length is also positively correlated with the temperature, although this correlation is weaker than the GC and temperature correlation. These correlations were not due to phylogenetic history, because they still persisted when the effect of phylogenetic non-independence was excluded. These findings support the argument that GC content of rRNA is an adaptation to the growth temperature. We further found that rRNA loop regions show a very constant base composition not affected by either differences in the optimal growth temperature or variations in the average GC content of the genome. This provides evidence that rRNA loops are subject to strong selective constraints, possibly in maintaining tertiary interactions within the single stranded regions. Therefore, it appears that the stems and loops of rRNA genes in a thermophilic species are subject to two kinds of selective pressures: (i) temperature-dependent selection for higher GC content and greater length in the stems, and (ii) selective pressure for the maintenance of tertiary

interactions in the loop regions. The latter selective pressure may prevent the overall rRNA becoming too high in GC content, thus preventing the formation of such unwanted structures like double stranded RNA.

Consistent with previous findings (e.g., Galtier & Lobry, 1997; Hurst & Merchant, 2001), we found that GC contents of whole genomes are not directly affected by growth temperature. This is against the selectionist view of thermal adaptation at the genomic level, recently renewed by the claim that there is a positive correlation between genomic GC and optimal growth temperature in prokaryotes when the relationship is considered within the same family (Musto et al, 2004). Indeed, for most bacteria, the range from minimum to maximum growth temperature is 30 °C. On the other hand, for the same temperature optimum (e.g., 37 °C), genomic GC content can be from 23.7% in *Mycoplasma bovoculi* to 69.5% in *Pseudomonas pseudomallei*. Bacterial genomic GC content seems related to genome size (a positive correlation between the two traits; Moran, 2002; Bastolla et al., 2004) and their life styles such as oxygen requirement (aerobic bacteria have a high GC than anaerobic ones; Naya et al., 2002; Lobry, 2004), habitat (host-associated bacteria have a lower GC content; Woolfit & Bromham, 2003; Rocha & Danchin, 2002) and salinity (halophiles have a higher GC content). A multiple regression of genomic GC content on genome size, growth temperature and the other life styles showed all the factors but the growth temperature are significant, indicating the environmental temperature plays a minor role in determining genomic GC content (H-C. Wang, unpublished result). These results indicate that different species employ different strategies to survive in high temperature.

In line with little correlation between genomic GC and the optimal growth temperature, the GC content of protein coding genes, which make the majority of a bacterial genome, is not correlated with growth temperature (Lambros, Mortimer & Forsdyke, 2003), nor is the GC3 of the protein genes (Hurst & Merchant, 2001). But protein coding genes can achieve whatever stability is necessary at the DNA level by other means, such as association of polyamines and relaxation of DNA supercoiling (Dalgaard & Garrett, 1993; Forsdyke & Bell, 2004), increased frequency of purines and polypurine tracts (Lambros, Mortimer & Forsdyke, 2003; Paz et al., 2004; more on Section 6.3.2). Furthermore, hyperthermophiles were found to have some unique proteins

(e.g., reverse gyrase and a putative chaperone) that are not present in any other organisms and some proteins (e.g., fructose 1,6-bisphosphase and RecB family exonuclease) are enriched in hyperthermophiles (Forterre, 2002; Makarova, Wolf & Koonin, 2003). These demonstrate that selection at high temperature involves many molecular processes simultaneously and genomic GC may not have to be elevated. Figure 6.1 summarizes the current understanding of genomic, transcriptomic and proteomic adaptations to growth at high temperature for prokaryotes.

Selective force		**Molecular level**		**Selective effect**
Selection for	→	Genome	→	No change in GC content (?)
growth at	→	Transcriptome	→	Double-stranded regions (e.g. rRNA stems) GC-rich; single-stranded regions (e.g., rRNA loops and mRNA) purine-rich.
high temperature	→	Proteome	→	Increase of charged residues; reduction of thermolabile residues; decrease in length; unique proteins (reverse gyrase, chaperone) and enrichment of certain enzymes.

Fig. 6.1 Molecular response to thermal adaptation (adapted from Hickey & Singer, 2004).

6.2 Effects of nucleotide bias on codon usage and protein evolution

The comparative analyses of sequences of homologous genes and proteins between rice and *Arabidopsis* indicated that genomic nucleotide bias causes amino acid composition, rather than to the opposite. This supports the neutralist view that directional mutational pressure would cause protein evolution and argues against the selectionist view that protein sequence is unrelated to genomic GC content. We showed that the mutational bias-driven protein evolution is in a predictable way. High GC rice genes encode proteins composed of high frequency of GC-rich codon encoded amino acids, i.e., glycine, alanine, arginine and proline. Low GC rice genes and *Arabidopsis* genes encode proteins

composed of high frequency of AT-rich codon encoded amino acids, i.e., phenylalanine, tyrosine, methionine, isoleucine, asparagines and lysine. Furthermore, rice genes have a gradient in GC content along the sequence with GC being higher at the 5-prime end. Consequently, the frequencies of G, A, R, and P amino acids are higher at the amino acid terminal of rice proteins.

Rates of synonymous nucleotide substitutions between high GC rice genes and their *Arabidopsis* homologs are on average higher than that between low GC rice genes and their *Arabidopsis* homologs, while the average rates of nonsynonymous substitutions between the *Arabidopsis* homologs and the high GC rice genes and low GC rice genes, respectively, are similar. These suggest mutational bias at the nucleotide level, rather than functional selection at the protein level, primarily determines the evolution of high-GC rice genes. These findings in multicellular higher plants, together with previous work on human proteins (Collins & Jukes, 1993), prokaryotic and yeast nuclear genes (Lobry, 1997; Gu, Hewett-Emmett & Li, 1998; Singer & Hickey, 2000) and mitochondrial genes (Foster, Jermiin & Hickey, 1997), provide pervasive evidence that directional mutational bias affects protein composition in a predictable way.

Nucleotide bias also has a direct effect on synonymous codon usage. We have shown that rice synonymous codon usage is primarily dictated by GC content of the gene. Like rice genes having a wide range of GC content, their patterns of codon usage also show large variations. High GC rice genes have more biased codon usage, which is not only distinct from their *Arabidopsis* homologs but also from low GC rice genes. Low GC rice genes have a codon usage pattern similar to that of their *Arabidopsis* homologs, for instance, they both have a preference for uridine and adenosine at the third codon position, while high GC rice genes have a preference for cytidine and guanosine. Part of the GC3s variation in rice genes is attributed to regional nucleotide compositional bias. Gene length also affects codon usage with shorter genes having higher codon bias. Translational selection may not play a major role in rice codon usage, as Nc-GC3 plot and codon adaptation index indicated ribosomal protein genes did not show a high codon bias. This is also supported by the evidence that a positive, rather than negative, correlation between synonymous substitution rates and codon bias.

6.3 Future directions

6.3.1 GC content, isochores and protein evolution in vertebrates

Given our current results showing the effects of nucleotide bias on protein evolution in prokaryotes and plants, one logical extension is to investigate this effect in vertebrate species, especially human. Although the relationship between GC content and human protein composition has been studied before (Collins & Jukes, 1993), and revealed the same trend as we observed here in rice and *Arabidopsis*, a more thorough analysis of the GC bias effect involving large scale comparative studies of homologous proteins between human and other vertebrates, e.g., pufferfish (*Fugu rubripes*) and mouse, will give us insights about the compositional structure and protein evolution patterns in vertebrates. The human and mouse genomes cover a very broad compositional spectrum and can be separated into fragments of more than 200 kb that are homogenous in GC content, called isochores. It will be interesting to compare amino acid composition and codon usage between those high, median and low GC human isochores and with their mouse homologs. This will help us to understand the evolution of isochore formation and the forces responsible for their maintenance in homeotherm vertebrates, which is still a matter of heated debate even though the complete human genome was sequenced four years ago (Eyre-Walker & Hurst, 2001).

Some recent studies of GC isochore evolution in mammalian genomes have led to the proposal that biased gene conversion (BGC) is the process responsible for their formation (Galtier et al., 2001). The key point is the finding that DNA recombination has a positive correlation with GC content (Fullerton, Carvalho & Clark, 2001; Meunier & Duret, 2004), so DNA segments high in GC have higher recombination rate - leading to biased gene conversion which, in turn, results in an even higher GC content. Evidence supporting this hypothesis includes studies of genes that are subject to concerted evolution, such as ribosomal operons, transfer RNAs, and histones - all of which are GC rich. In addition, genes in recombination hotspots, such as genes in the pseudoautosomal region (PAR) of the X and Y chromosomes have much higher GC3 than genes located in autosomes and non-PAR regions of the sex chromosomes (Galtier et al., 2001). Mammalian histone-coding genes have presumably undergone gene conversion resulting

in a higher GC content than in histone genes that have not undergone gene conversion (Galtier, 2003). To further test the causal relationship between recombination and GC content and to understand the role of biased gene conversion in isochore evolution, it would be interesting to examine allelic polymorphism patterns in inbreeding species, populations of which are homozygotes so that BGC, however strong, has no effect on GC content (Galtier et al., 2001). Inbreeding species are rare in vertebrates but common in plants. The genomes of Gramineae are quite similar to mammalian genomes with regards the distribution of both GC content and genes (Carels & Bernardi, 2000). It will be interesting to compare sequence polymorphism patterns in inbreeding and outbreeding Gramineae species to understand GC content evolution in higher plants and to help delineate the role of BGC in mammalian GC evolution.

6.3.2 Nucleotide bias, thermal adaptation and other forms of selection

It was pointed out (Zhang & Zhang, 1991; Zhang & Chou, 1994) that base composition of a DNA can be uniquely represented by three parameters x, y and z:

$$\begin{cases} x = (A + G) - (C + T) \\ y = (A + C) - (G + T) \\ z = (A + T) - (C + G) \end{cases} \qquad \textbf{(6.1)}$$

where x is the difference of purines and pyrimidines (i.e., measuring the ratio of purine to pyrimidine R/Y), y is the difference of 6-amino bases and 6-oxo derivative (keto) bases, and z is the difference of weak and strong hydrogen bond bases. We have specifically focused on GC content, i.e., the z parameter, in thermal adaptation of rRNA genes and established that rRNA genes (and especially rRNA stems) have a higher GC content in thermophilic bacteria and archaea. For rRNA loops we found A and G are particularly higher in both mesophilic and thermophilc species (Table 2.3B and C; Fig. 2.2), indicating a high value of the x parameter. The x parameter is also called purine loading meaning the excess of purines in a sequence (Lao & Forsdyke, 2000). Purine loading has also been demonstrated in protein coding genes, i.e., mRNAs, and found to relate with transcription direction of the gene (Szybalski et al., 1966; Bell & Forsdyke, 1999).

Although GC contents of the genome and protein coding genes are not correlated with optimal temperature in prokaryotes, purine loading is found positively correlated with temperature (Lambros, Mortimer & Forsdyke, 2003; Paz et al., 2004). But these studies did not take into account phylogenetic relatedness among the species. Since the correlation is generally weak (Lambros, Mortimer & Forsdyke, 2003), it would be interesting to analyse it in a phylogeny independent context, for example doing phylogenetic independent contrast analyses and comparing purine loading of mesophilic and thermophilic species within the same genus. Since purine loading is related to transcription direction (i.e., it is the RNA-synonymous strand that is purine-loaded), it is important to distinguish genes transcribed to the right of the promoters and those transcribed to the left when doing these analyses (Forsdyke & Bell, 2004).

Furthermore, studies have found there are changes in amino acid composition in proteins and enzymes between mesophilic and thermophilic prokaryotic species (Deckert et al., 1998; Kawarabayasi et al., 1999; Haney et al., 1999; Singer & Hickey, 2003; Hickey & Singer, 2004). It is interesting to compare nucleotide bias in this protein composition change. It is also important to analyze molecular mechanisms of protein thermostability, which will be especially useful in engineering thermal stable enzymes that have commercial applications in molecular biology and biotechnology (Vieille & Zeikus, 2001). Some recent studies compared phylogenetic pattern of orthologous proteins of (hyper)thermophilic and mesophilic species and identified which proteins are unique or enriched in the (hyper)thermophiles to understand potential genomic determinants of (hyper)thermophily (Forterre, 2002; Makarova, Wolf & Koonin, 2003). This approach will become even more useful when complete genome sequence of more (hyper)thermophiles are available.

Some bacteria are subject to other selective forces such as high salt or highly alkaline environments, or other extreme conditions like acidity and pressure, or UV radiation, or lack or low oxygen and host dependence (endosymbioticity). Recent studies have found genomic GC content is lower in anaerobic bacteria than in aerobacteria (Naya et al., 2002; Lobry, 2004). This has linked for the first time a genome structure (i.e., GC content) and bacterial life history trait, suggesting of an impact of genomic GC content on bacterial fitness. It will be interesting to see whether GC content of structural RNAs is

also higher in aerobiosis. It is also useful to study amino acid composition of proteins in this process and two recent studies have found aerobic bacteria tend to use amino acids that are inexpensive in terms of metabolic cost (Akashi & Gojobori. 2002; Lobry, 2004).

Genomic adaptation to host dependence (endosymbioticity) of microorganisms is also of interest. Endosymbiotic bacteria and fungi are subject to severe population bottlenecks at each host generation. A recent study shows that while there is no significant difference in GC content of 16S rRNA genes between endosymbiotic bacteria and fungi and their non-endosymbiont relatives, the average overall genomic GC contents are significantly different between the two kinds of microorganisms, with endosymbionts being GC-poorer (i.e., AT richer) (Woolfit & Bromham, 2003). Another study also found obligatory pathogenic bacteria have a higher average genomic A+T content (62% A+T) than free-living or facultively pathogenic bacteria (51% A+T) (Rocha & Danchin, 2002). These two studies did not analyse GC content difference in the three codon positions between the endosymbionts and nonendosymbionts, but this (especially comparing GC3) would be most worthwhile to do in order to understand the mechanisms that bring about the nucleotide bias. Furthermore, since GC content is negatively related to purine loading (Lambros, Mortimer & Forsdyke, 2003) it would be also useful to analyse purine loading in these cases.

6.3.3 Other perspectives

So far we have not mentioned parameter y (i.e., **A+C-G-T**) in equations 6.1. This bias often reflects an asymmetry in the mutational bias in the leading and lagging strands of replication and transcription (Lobry, 1996) and it has been found in several prokaryotes (McLean, Wolfe & Devine, 1998). In *Borrelia burgdorferi* and *Chlamydia trachomatis* this GT-bias difference in the leading and lagging strands has been identified as the most important source of variation in codon usage (McInerney, 1998; Romero, Zavala & Musto, 2000). It would be useful to analyse GT-bias with regards transcription directions in analyzing rRNA and mRNA thermal adaptations.

Graphical display of the three nucleotide bias parameters shown in Equations 6.1 on a three dimensional space is called z curving (Zhang & Zhang, 1991; Zhang & Chou, 1994) and it has been used successfully to identify protein coding sequences (Zhang &

Wang, 2000; Guo, Ou & Zhang, 2003). Similarly, the genomic GC bias has been used in identifying structural RNA genes in AT-rich hyperthermophiles (Rivas & Eddy, 2000; Carter et al., 2001; Klein, Misulovin & Eddy, 2002). Since the *z*-curve method simultaneously displays the three nucleotide bias parameters and since thermal adaptation involving at least two of the parameters (GC content and purine loading) it would be of interest in analyzing *z*-curve difference between mesophilic and thermophilic organisms, which can also be used in identifying structural RNAs. As described above, it would also be worthwhile to compare the z-curve patterns between endosymbionts and non-endosymbionts.

The comparative analyses of rice and *Arabidopsis* genomes conducted in this project have shown that mutational bias causes both codon usage bias and amino acid bias. As shown in Chapter 4.4.4 (Fig. 4.10) the nucleotide bias also affects protein similarity scores. The phylogenetic implications of DNA-driven amino acid compositional bias has previously been recognized as a problem in phylogenetic analysis (Foster & Hickey, 1999; Foster, 2004) and needs to be addressed when designing phylogenetic software.

The G+C content of DNA has many implications that are not addressed in this project. For instance, biased base composition and codon usage may be used as an index of horizontal gene transfer (Medigue et al. 1991; Lawrence & Ochman 1998; Wang et al., 2001; but see also Koski, Morton & Golding, 2001; Wang, 2001). It is important to model compositional change of horizontally transferred genes (Lawrence & Ochman 1997), as this, after all, will affect nucleotide composition of the host genome. The human and mouse genomes have a lot of gene order breakpoints (Nadeau & Taylor, 1984; Eichler & Sankoff, 2003) and extensive breakpoint reuse (Pevzner & Teslet, 2003). It will be interesting to analyse nucleotide bias in this process.

Finally, GC contents (and composition of individual nucleotides) that we have analysed in this thesis are in most cases the average percentage of G+C (or individual nucleotides) relative to the total number of nucleotides in the sequence or the whole genome of interest. There sometimes is a need to analyse GC content along a sequence or even a whole genome so that the positional distribution of GC content can be viewed. One such plot of genomic GC content of *M. jannachii* is shown in Figure 2.5 (see

Chapter 2.5). This kind of plots helps in finding positional GC content variations, which may lead to finding substantial structures. In Figure 2.5 for instance some structural RNA genes may be identified by simply viewing the graph. Figure 4.7 (see Chapter 4.4.3) presents another case where a gradient in GC content can be seen in a GC plot. In these cases the GC content is often calculated and smoothed over a sliding window of sequences. It is important to choose the proper length of a window, but this may be difficult in practice. Indeed, such a difficulty arose when analyzing the isochore structure of human genome (The genome sequencing consortium, 2001; Clay & Bernardi, 2001). The delineating of relative homogeneous GC domains will be an interesting topic for mathematical genomics (Li, 2001; Wan & Wootton, 2000) and will be useful for analyzing GC content variation and genome evolution.

References

Akashi H, Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc Natl Acad Sci USA* 99:3695-3700.

Altschul SF, Gish, W, Miller W, Myers EW and Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403-410.

Alvarez-Valin F, Jabbari K, Carels N and Bernardi G. 1999. Synonymous and nonsynonymous substitutions in genes from Gramineae: intragenic correlations. *J Mol Evol* 49:330-342.

Alvarez-Valin F, Lamolle G and Bernardi G. 2002. Isochores, GC3 and mutation biases in the human genome. *Gene* 300:161-168.

Badger, J. 1999. *Exploration of Microbial Genomic Sequences via Comparative Analysis.* Ph.D. thesis, University of Illinois at Urbana-Champaign.

Bairoch A and Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28: 45-48.

Ban N, Nissen P, Hansen J, Moore PB and Steitz TA. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* 289, 905–920.

Bastolla U, Moya A, Viguera E and van Ham RCHJ. 2004. Genomic determinants of protein folding thermodynamics in prokaryotic organisms. *J Mol Biol* 343:1451-1466.

Bell SJ and Forsdyke DR. 1999. Deviations from Chargaff's second parity rule correlate with direction of transcription. *J theor Biol* 197:63-76.

Belle EM, Smith N and Eyre-Walker A. 2002. Analysis of the phylogenetic distribution of isochores in vertebrates and a test of the thermal stability hypothesis. *J Mol Evol* 55:356-363.

Belozersky AN and Spirin AS. 1958. A correlation between the compositions of deoxyribonucleic and ribonucleic acids. *Nature* 182:111.

Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* 241:3-17.

Bernardi G and Bernardi G. 1986. Compositional constraints and genome evolution. *J Mol Evol* 24:1-11.

Bernardi G, Olfsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M and Rodier F. 1985. The mosaic genome of warm-blooded vertebrates. *Science* 228:953–958.

Bill CA, Duran WA, Miselis NR and Nickoloff JA. 1998. Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese Hamster Ovary cells: competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches. *Genetics* 149:1935-1943.

Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol* 19:1181-1197.

Brochier C, Forterre P and Gribaldo S. 2004. Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus kandleri paradox. *Genome Biol* 5:R17

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG., Blake JA, FitzGerald LM, Clayton RA, Gocayne JD et al. 1996. Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science* 273:1058–1073.

Cambillau C and Claverie J-M. 2000. Structural and genomic correlates of hyperthermostability. *J Biol Chem* 275:32383-32386.

Carbone A, Zinovyev A and Kepes F. 2003. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* 19:2005-2015.

Carels N and Bernardi G. 2000. Two classes of genes in plants. *Genetics* 154:1819-1825.

Carels N, Hatey P, Jabbari K, Bernardi G. 1998. Compositional properties of homologous coding sequences from plants. *J. Mol. Evol.* 46:45-53.

Carter RJ, Dubchak I and Holbrook SR. 2001. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res* 29:3928-3938.

Chargaff E. 1950. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* 6:201-209.

Chargaff E. 1979. How genetics got a chemical education. *Annals New York Acad Sci* 325:345-360.

Chen L-L and Zhang C-T. 2003. Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages. *Biochem Biophys Res Commun* 306:310-317.

Clay O and Bernardi G. 2001. Compositional heterogeneity within and among isochores in mammalian genomes. II. Some general comments. *Gene* 276:25-31.

Collins DW and Jukes TH. 1993. Relationship between G + C in silent sites of codons and amino acid composition of human proteins. *J Mol Evol* 36:201-213.

Dalgaard JZ and Garrett RA. 1993. Archaeal hyperthermophile genes. In Kates M, Kushner DJ and Matheson AT (eds), *The Biochemistry of Archaea (Archaebacteria).* Elsevier, Amsterdam, pp. 535–563.

De Rijk P. 1995. *Optimisation of a database for ribosomal RNA structure and application in structural and evolutionary research.* PhD thesis, University of Antwerp, Antwerp, Belgium.

Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M et al. 1998. The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. *Nature* 392:353–358.

Drake JW, Charlesworth B, Charlesworth D and Crow JF. 1998. Rates of Spontaneous Mutation. *Genetics* 148:1667-1686.

Duret L and Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci USA* 96:4482-4487.

Duret L, Mouciroud D and Gautier C. 1995. Statistical analysis of vertebrate seqeunces reveals that long genes are scarce in GC-rich isochores. *J Mol Evol* 40:308-317.

Duret L, Semon M, Piganeau G, Mouchiroud D and Galtier N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* 162:1837-1847.

Eichler EE and Sankoff D. 2003. Structural dynamics of eukaryotic chromosome evolution. *Science* 301:793-797.

Eyre-Walker A and Hurst LD. 2001. The evolution of isochores. *Nature Reviews: Genetics* 2:549-555.

Felsenstein J. 1985. Phylogeny and the comparative method. *Am Naturalist* 125:1-15.

Felsenstein J. 2004. *Inferring Phylogenies*. Sinauer Associates, Inc. Sunderland.

Felsenstein, J. 2004b. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Fennoy SL and Bailey-serres J. 1993. Synonymous codon usage in Zea mays L. nuclear genes is varied by levels of C and G-ending codons. *Nucleic Acids Res* 21:5294-5300.

Forsdyke DR and Bell SJ. 2004. Purine-loading, stem-loops, and Chargaff's second parity rule. *Applied Bioinfo* 3:3-8.

Forsdyke DR and Mortimer JR. 2000. Chargaff's legacy. *Gene* 261:127-137.

Foster PG. 1997. *Phylogenetic Implications of the Effect of Nucleotide Bias on Amino Acid Composition.* PhD thesis. University of Ottwa.

Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol* 53:485-95.

Foster PG and Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol* 48:284-290.

Foster PG, Jermiin LS and Hickey DA. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol* 44:282-288.

Forterre P. 2002. A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. *Trends Genet* 18:236-237.

Fullerton SM, Carvalho AB and Clark AG. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol Biol Evol* 18:1139-1142.

Gautier C. 2000. Compositional bias in DNA. *Curr Opin Genet Dev* 10:656-661.

Galtier N. 2003. Gene conversion drives GC content evolution in mammalian histones. *Trends Genet* 19:65-68.

Galtier N and Lobry JR. 1997. Relationships between genomic GC content, RNA secondary structures and optimal growth temperature in prokaryotes. *J Mol Evol* 44:632-636.

Galtier N, Piganeau G, Mouchiroud D and Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159:907-911.

Goff, SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* 296:92-100.

Grafen A and Ridley M. 1996. 'Statistical tests for discrete cross-species data'. *J theor Biol* 183:255-267.

Grantham R., Gautier C, Gouy M, Mercier R and Pave A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res* **8**:r49–r62.

Greenacre, M. J. 1984. *Theory and Applications of Correspondence Analysis.* Academic Press, London.

Grocock RJ and Sharp PM. 2002. Synonymous codon usage in Pseudomonas aeruginosa PA01. *Gene* 289:131-139.

Gu X, Hewett-Emmett D and Li WH. 1998. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica* 102-103: 383-391.

Guo FB, Ou HY and Zhang CT. 2003. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res* 31:1780-1789.

Gutell RR, Cannone JJ, Sheng Z, Du Y and Serra MJ. 2000. A story: unpaired adenosine bases in ribosomal RNAs. *J Mol Biol* 304:335-354

Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR and Olsen GJ. 1999. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species. *Proc Natl Acad Sci USA* 96:3578-3583.

Harvey PH and Pagel MD. 1991. *The Comparative Method in Evolutionary Biology.* Oxford University Press, Oxford.

Harvey PH and Rambaut A. 1998. Phylogenetic extinction rates and comparative methodology. *Proc Roy Soc Lond B* 265:1691-1696.

Hasegawa M and Hashimoto T. 1993. Ribosomal RNA trees misleading? *Nature* 361:23.

Hickey DA, Bally-Cuif L, Abukashawa S, Payant V and Benkel BF. 1991. Concerted evolution of duplicated protein-coding genes in Drosophila. *Proc Natl Acad Sci USA* 88: 1611-1615.

Hickey DA and Singer GAC. 2004. Genomic and protemic adaptations to growth at high temperature. Genome Biol 5:117

Hurst LD and Merchant AR. 2001. High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proc R Soc Lond B* 268: 493-497.

Jukes TH and Cantor CR. 1969. Evolution of protein molecules. pp. 21-132. In HN Munro (ed.) *Mammalian Protein Metabolism.* Academic Press, New York.

Karlin S, Campbell AM and Mrazek J. 1998. Comparative DNA analysis across diverse genomes. *Annu Rev Genet* 32:185-225.

Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jin-no K, Takahashi M, Sekine M, Baba S, Ankai A et al. 1999. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, Aeropyrum pernix K1. *DNA Res* 6:83–101.

Kawarabayasi Y, Sawada,M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A et al. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3. *DNA Res* 5, 55–76.

Kerr AR, Peden JF and Sharp PM. 1997. Systematic base composition variation around the genome of Mycoplasma genitalium, but not Mycoplasma pneumoniae. *Mol Microbiol* 25:1177-1179.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol.* 16:111-120.

Klein RJ, Misulovin Z, Eddy SR. 2002. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc Natl Acad Sci USA* 99:7542-7547.

Klenk HP, Clayton RA, Tomb J, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. *Nature* 390: 364–370.

Koski LB, Morton RA, Golding GB. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol* 18:404-412.

Kunkel TA and Alenxander PS. 1986. The base substitution fidelity of eukaryotic DNA polymerases. *J Biol Chem* 261:160-166.

Kreil DP and Ouzounis CA. 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res* 29:1608-1615.

Lambros RJ, Mortimer JR and Forsdyke DR. 2003. Optimum growth temperature and the base composition of open reading frames in prokaryotes. *Extremophiles* 7:443-450.

Lao PJ and Forsdyke DR. 2000. Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Res* 10:228-236.

Lawrence JG and Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 44:383–397.

Lawrence JG and Ochman H. 1998. Molecular archaeology of Escherichia coli genome. *Proc Natl Acad Sci USA* 95:9413–9417.

Lee KY, Wahl R, and Barbu E. 1956. Content of purine and pyrimidine base in desoxyribonucleic acid of bacteria. *Ann Inst Pasteur* 91:212.

Li W. 2001. Delineating relative homogeneous G+C domains in DNA sequences. *Gene* 276:57-72.

Li WH. 1997. *Molecular Evolution*. Sinauer Associates, Inc. Sunderland.

Li WH and Graur D. 1991. *Fundamentals of Molecular Evolution*. Sinauer Associates Inc, Sunderland.

Liu Q, Feng Y, Zhao X, Dong H and Xue Q. 2004. Synonymous codon usage bias in Oryza sativa. *Plant Science* 167:101-105.

Lobry JR. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* 13:660-665.

Lobry JR. 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene* 205:309-316.

Lobry JR. 2004. Life history traits and genomic structure: aerobiosis and G+C content in bacteria. *Lecture Notes in Computer Science* 3039:79-686.

Lobry JR and Sueoka N. 2002. Asymmetric directional mutation pressures in bacteria. *Genome Biol* 3(10):RESEARCH0058.

Loomis WF and Smith DW. 1990. Molecular phylogeny of Dictyostelium discoideum by protein sequence comparison. *Proc Natl Acad Sci USA* 87:9093-9097.

Makarova KS, Wolf YI and Koonin EV. 2003. Potential genomic determinants of hyperthermophily. *Trends Genet* 19:172-176.

Martins EP and Garland T Jr. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution* 45:534-557.

Martins EP and Housworth EA. 2002. Phylogeny shape and the phylogenetic comparative method. *Syst Biol* 51: 873-880.

Matassi G, Montero LM, Salinas J, Bernardi G. 1989. The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants. *Nucleic Acids Res* 17:5273-5290.

Mathe C, Peresetsky A, Dehais P, Van Montagu M and Rouze P. 1999. Classification of Arabidopsis thaliana gene sequences: clustering of coding sequences into two groups according to codon usage improves gene prediction. *J Mol Biol.* 285:1977-1991.

McInerney JO. 1998. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. *Proc Natl Acad Sci USA*. 95:10698-10703.

McLean MJ, Wolfe KH and Devine KM. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol* 47:691-696.

Médigue C, Rouxel T, Vigier P, Henaut A and Danchin A. 1991. Evidence of horizontal gene transfer in Escherichia coli speciation. *J Mol Biol* 222:851–856.

Meunier J and Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol* 21:984-990.

Milton JS. 1992. *Statistical Methods in the Biological and Health Sciences*. McGraw-Hill, Inc., New York, p.368.

Mooers AO and Holmes EC. 2000. The evolution of base composition and phylogenetic inference. *Trends Ecol Evol* 15:365-369.

Moran NA. 2002. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 10:583-586.

Morton BR. 1999. Strand asymmetry and codon usage bias in the chloroplast genome of Euglena gracilis. *Proc Natl Acad Sci USA* 96:5123-5128.

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F and Bernardi G. 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett* 573:73-77.

Muto A and Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci USA* 84, 166-169.

Nadeau JH and Taylor BA. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci USA* 81:814-818.

Nagaswamy U, Larios-Sanz M, Hury J, Collins S, Zhang Z, Zhao Q and Fox GE. 2002. NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acids Res* 30:395–397.

Nakachi Y, Hayakawa T, Oota H, Sumiyama K, Wang L, Ueda S. 1997. Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. *Mol Biol Evol* 14:1042-1049.

Nakashima H, Fukuchi S and Nishikawa K. 2003. Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. *J Biochem* 133:507-513.

Naya H, Romero H, Zavala A, Alvarez B and Musto H. 2002. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J Mol Evol* 55:260-264.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC., Ketchum KA et al. 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* 399:323–329.

Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol* 19:1390-1394.

Oliver JL and Marin A. 1996. A relationship between GC content and coding-sequence length. *J Mol Evol* 43:216-223.

Paz A, Mester D, Baca I, Nevo E and Korol A. 2004. Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. *Proc Natl Acad Sci USA* 101:2951-2956.

Peden JF. 1999. *Analysis of Codon Usage*. Ph.D. Thesis, University of Nottingham.

Pevzner P and Tesler G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci USA* 100:7672-7677.

Powell JR, Moriyama EN. 1997. Evolution of codon usage bias in Drosophila. *Proc Natl Acad Sci USA* 94:7784-7790.

Rice P Longden I and Bleasby A. 2000. EMBOSS: The European molecular biology open software suite. *Trends Genetics* 16:276-277.

Rivas E and Eddy SR. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* 16:583-605.

Rocha EP and Danchin A. 2002. Base composition bias might result from competition for metabolic resources. *Trends Genet* 18:291-294

Rolfe R and Meselson M. 1959. The relative homogeneity of microbial DNA. *Proc Natl Acad Sci USA* 45:1039–1043.

Romero H, Zavala A and Musto H. 2000. Codon usage in Chlamydia trachomatis is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res* 28:2084-2090.

Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, Mewes HW, Frishman D, Stocker S, Lupas AN and Baumeister W. 2000. The genome sequence of the thermoacidophilic scavenger Thermoplasma acidophilum. *Nature* 407:508–513.

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y, et al. 2002. The genome sequence and structure of rice chromosome 1. *Nature* 420:312-316.

Shannon CE. 1948. A mathematical theory of communication. *Bell System Tech J* 27:379–423, 623–656.

Sharp PM and Li WH. 1987a. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15:1281-1295.

Sharp PM and Li WH. 1987b. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* 4:222-230.

She Q, Singh RK, Confalonieri F, Zivanovic Y, Allard G, Awayez MJ, Chan Weiher CC, Clausen IG., Curtis,BA, De Moors A et al. 2001. The complete genome of the crenarchaeon Sulfolobus solfataricus P2. *Proc Natl Acad Sci USA* 98:7835–7840.

Singer CE and Ames BN. 1970. Sunlight ultraviolet and bacterial DNA base ratios. *Science* 170:822-826.

Singer GAC. 2002. *Non-random Neutral Evolution*. PhD thesis. University of Ottawa.

Singer GAC and Hickey DA. 2000. Nucleotide bias causes genomewide bias in the amino acid composition of proteins. *Mol Biol Evol*, 17:1581-1588.

Singer GAC and Hickey DA. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* 317:39-47

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H-M, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K et al. 1997. Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. *J Bacteriol*, 179:7135–7155.

SPSS Science. 2000. *SYSTAT version 10*, Chicago.

Stoesser G, Baker W, van den Broek A, Camon E, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V, et al 2002. The EMBL nucleotide sequence database. *Nucleic Acids Res* 30:21-26.

Sueoka N. 1961a. Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harb Symp Quant Biol* 26:35-43.

Sueoka N. 1961b. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc Natl Acad Sci USA* 47:1141-1149.

Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci USA* 48:582-592.

Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci USA* 85:2653-2657.

Sueoka N. 1999. Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C. *J Mol Evol* 49:49-62.

Sueoka N, Marmur J and Doty P. 1959. Dependence of the density of deoxyribonucleic acids on guanine-cytosine content. *Nature* 183:1429-1431.

Suzuki H, Saito R and Tomita M. 2004. The 'weighted sum of relative entropy': a new index for synonymous codon usage bias. *Gene* 335:19-23.

Szybalski W, Kubinski H and Sheldrick P. 1966. Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis. *Cold Spring Harb Symp Quant Biol* 31:123-127.

Tamura K and Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10: 512-526.

Tamura K and Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol* 19:1727-1736.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowing plant Arabidopsis thaliana. *Nature* 408:796-815.

The Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860-921.

Thompson JD, Higgins DG and Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673-4680.

Tillier ER and Collins RA. 2000. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J Mol Evol* 50:249-257.

Urrutia AO and Hurst LD. 2003. The signature of selection mediated by expression on human genes. *Genome Res* 13:2260-2264.

Van de Peer Y, De Rijk P, Wuyts J, Winkelmans T and De Wachter R. 2000. The European small subunit ribosomal RNA database. *Nucleic Acids Res* 28:175-176.

Vieille C and Zeikus GJ. 2001. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev* 65:1-43.

Wada A and Suyama A. 1986. Local stability of DNA and RNA secondary structure and its relation to biological functions. *Prog Biophys Mol Biol* 47:113-157.

Wan H and Wootton JC. 2000. A global compositional complexity measure for biological sequences: AT-rich and GC-rich genomes encode less complex proteins. *Comput Chem* 24:71-94.

Wang B. 2001. Limitations of compositional approach to identifying horizontally transferred genes. *J Mol Evol* 53:244-250.

Wang HC, Badger J, Kearney P and Li M. 2001. Analysis of codon usage patterns of bacterial genomes using the self-organizing map. *Mol Biol Evol* 18:792-800.

Wang HC and Hickey DA. 2002. Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes. *Nucleic Acids Res* 30:2501-2507.

Wang HC, Singer GAC and Hickey DA. 2004. Mutational bias affects protein evolution in flowing plants. *Mol Biol Evol* 21:90-96.

Wang X, Shi X and Hao B. 2002. The transfer RNA genes in Oryza sativa L. ssp. indica. *Science in China* 45:504-511.

Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, Teytelman L, Schmidt S, Zhao W, Cartinhour S et al. 2002. Gramene: a resource for comparative grass genomics. *Nucleic Acids Res* 30:103-105.

Watson JD and Crick FHC. 1953. Molecular structure of nucleic acids. *Nature* 171:737-738.

Wilquet V and Van de Casteele M. 1999. The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. *Res Microbiol* 150:21-32.

Wimberly BT, Brodersen DE, Clemons WM,Jr, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T and RamakrishnanV. 2000. Structure of the 30S ribosomal subunit. *Nature* 407:327–339.

Wong GK, Wang J, Tao L, Tan J, Zhang J, Passey DA and Yu J. 2002. Compositional gradients in Gramineae genes. *Genome Res* 12:851-856.

Woolfit M and Bromham L. 2003. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Mol Biol Evol* 20:1545-1555.

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

Wright S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc 6th Internatl Congress Genetics* 1:356-366.

Wuyts J, Van de Peer Y, Winkelmans T and De Wachter R. 2002. The European database on small subunit ribosomal RNA. *Nucleic Acids Res* 30:183-185.

Xia X. 1996. Maximizing transcription efficiency causes codon usage bias. *Genetics* 144:1309-1320.

Xia X. 2000. *Data Analysis in Molecular Biology and Evolution.* Kluwer Academic Publishers.

Xia X, Xie Z and Li WH. 2003. Effects of GC content and mutational pressure on the lengths of exons and coding sequences. *J Mol Evol* 56:362-370.

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X et al, 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* 296:79-92.

Zhang CT and Chou KC. 1994. A graphic approach to analyzing codon usage in 1562 Escherichia coli protein coding sequences. *J Mol Biol* 238:1-8.

Zhang CT and Wang J. 2000. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. *Nucleic Acids Res* 28:2804-2814.

Zhang CT and Zhang R. 1991. Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res* 19:6313-6317.