

T 162

49659

CENTRAL LIBRARY	
TEZPUR UNIVERSITY	
Accession No. <u>49659</u>	CENTRAL LIBRARY, T. U. ACC. NO. <u>T.162</u>
Date <u>14/9/11</u>	

**REFERENCE BOOK
NOT TO BE ISSUED
TEZPUR UNIVERSITY LIBRARY**

**A STATISTICAL STUDY ON THE NUCLEOTIDE
COMPOSITION OF BACTERIAL
CHROMOSOMES**

**A THESIS SUBMITTED IN PART FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY**

Mr. Bhesh Raj Powdel

Registration No. 002 of 2010



Department of Mathematical Sciences

School of Science and Technology

Tezpur University

Tezpur- 784028, Assam, India

Dedicated to the memory of my father

Late Mukti Nath Powdel



TEZPUR UNIVERSITY

This is to certify that the thesis entitled "A Statistical Study on the Nucleotide Composition of Bacterial Chromosomes" submitted to the School of Science & Technology, Tezpur University in part fulfilment for the award of the degree of Doctor of Philosophy in Mathematical Sciences is a record of research work carried out by Mr. Bhesh Raj Powdel under our supervision and guidance.

All help received by him from various sources have been duly acknowledged.

No part of this thesis has been submitted elsewhere for award of any other degree.

Signatures of


Supervisor: Prof. Munindra Borah


Co-Supervisor: Dr. Suendra Kumar Ray

Designation: Professor

Designation: Associate Professor

Department: Mathematical Sciences

Department: Molecular Biology &
Biotechnology

School: Science & Technology

Affiliation: Tezpur University

PREFACE

Deoxyribonucleic acid (DNA) is the carrier of genetic information in organisms. The structure of DNA was elucidated in 1953 by Watson and Crick. DNA is made up of two polynucleotide chains twisted around each other in a right handed fashion (right handed double helix). DNA is a heteropolymer composed of four different monomers A (adenine), C (cytosine), G (guanine) and T (thymine). The two strands run in opposite directions with respect to each other (antiparallel) and are held together by hydrogen bonds between complementary bases of W (A and T with two hydrogen bonds) and S (G and C with three hydrogen bonds) nucleotides. This base pairing in DNA has implication on its role in inheritance. The sequential arrangement of nucleotides is central to its function in organisms.

A bacterial genome is the collection of a bacterium's entire genetic information. Essentially, it determines how a bacterium looks and functions, both externally and internally. This genetic information is organized into genes, which are encoded in the organism's DNA. Those genes are further organized into chromosomes. All bacteria are haploid, and possess either one or more chromosomes.

The year 1977 was the beginning of DNA sequencing. The first bacterial genome to be sequenced was *Haemophilus influenza* in the year 1995. During the past 15 years of rapid developments in genomic and other molecular research technologies particularly in the field of genome sequencing projects and developments in information technologies have combined to produce a tremendous amount of information related to genomes. After 15 years, on the 30th June 2010, 1104 bacterial genome sequences are available in the DDBJ web site (www.gib.genes.nig.ac.jp). These developments have invited mathematical and computing approaches to the understanding of biological processes. A term Bio-informatics was used in the year 1979 to this inter disciplinary field of mathematics, biology and information technology. The sole aim of bioinformatics is to increase our understanding by analysing huge amount of biological data. Computationally intensive techniques are used to recognize patterns, data mining, algorithms and visualization of biological systems. Major research efforts in the field include sequence alignment, gene finding, drug designing, protein structure alignment, prediction of gene expression, modelling of evolution etc.

In the genomic era, our understandings on evolutionary aspects of microbial genomes have increased significantly. Some of the important findings relating to whole genome compositional studies are Chargaff's 2nd parity in chromosomes, strand specific mutational

bias, AT enrichment towards the terminus of bacterial chromosomes, and codon usage difference between the leading and lagging strands in chromosomes. The objectives of this PhD research are based on these findings and following are some descriptions of these evolutionary findings.

In a double stranded DNA, the complementary base pairing rule puts a constraint of equimolar frequencies of the complementary bases i.e. $f_A = f_T$ and $f_G = f_C$ where $f_A + f_T + f_G + f_C = 1$. The compositional similarity between complementary nucleotides in double stranded DNA is known as Chargaff's rule, which was discovered in 1950. The chemical composition of individual DNA strands was reported in 1968 also from Chargaff's laboratory for *Bacillus subtilis* chromosome and extended later to six more bacterial species. Chargaff and his colleagues observed the similarity between the abundance values of complementary nucleotides ($f_A \approx f_T$, $f_G \approx f_C$) within individual DNA strands of bacterial chromosomes, which was very surprising for them. In the post genomic era, the compositional similarity between complementary nucleotides is observed in chromosomes of bacteria, archaea and eukaryotes, which is now known as Chargaff's 2nd parity or intra-strand parity. Although large numbers of papers have been published citing works and discussions on intra-strand parity in the genomic era, scientists are yet to find all the factors responsible for such a universal phenomenon in the chromosomes.

Replication at each of the forks is an asymmetric process: on the leading strand (LeS) it proceeds continuously, whereas on the lagging strand (LaS) it proceeds discontinuously by the synthesis and joining of short Okazaki fragments. Under no bias between LeS and LaS with respect to mutation and selection, intra-strand parity or Chargaff's 2nd parity is likely to be observed in chromosomes, which is called as parity rule 2 (PR2). Any deviation from PR2 implies asymmetric substitution rates, different selective pressures in the two strands of DNA. The asymmetry during replication has been shown to affect differentially mutation/nucleotide-substitution rates as well as gene distributions between the LeS and the LaS. In most of the bacteria higher frequency of the keto nucleotides (G & T) is observed in the LeS in comparison to the LaS. In some other bacteria higher frequency of the purine nucleotides (G & A) is observed in the LeS in comparison to the LaS. This compositional asymmetry between the strands is known as strand specific mutational bias (SSMB) SSMB is observed in most of bacterial genomes analysed till date. SSMB has been used to predict origin and terminus of replication in bacterial chromosomes. SSMB is known to affect codon usage bias (CUB) in genomes. In some bacteria the effect of SSMB is so much that the codon

usage bias in genes is determined by their strand location rather than their expression. Though SSMB is known to affect CUB in organisms, its effect on highly expressed genes and weakly expressed genes whether same or different is yet to be studied in bacteria.

Non-random usage of synonymous codons, otherwise called as codon usage bias (CUB), is common in prokaryotes, eukaryotes and viruses. Patterns and degrees of CUB vary not only among different organisms, but also among genes in the same genome. CUB is affected by both mutation and selection pressures in organisms. The major challenge for molecular evolutionary biologists is to estimate the selection responsible for codon usage bias in a gene. The codon usage bias due to selection is termed as selected codon usage bias. The selection-mutation-drift (SMD) theory suggests that in weakly expressed genes codon usage bias is determined mainly by mutation whereas in highly expressed genes codon usage bias is determined mainly by selection. Sharp *et al.* (2005) introduced the population genetics-based model (Bulmer, 1991) for quantifying the extent to which selection has been effective on codon usage bias in an organism. They observed variable strength of selected codon usage bias among bacteria. Bacterium such as *Escherichia coli* with low SSMB was found with strong selected codon usage bias and bacterium such as *Borrelia burgdorferi* with high SSMB was found with weak selected codon usage bias. They had not directly compared the strength of selected codon usage bias with SSMB in genomes. So it is not clear whether the genomes with strong selected codon usage bias belong to both low and high SSMB or only to low SSMB group. One of the objectives in the present thesis is to study this phenomenon.

Objectives

The objective of this research is to do a statistical analysis of nucleotide composition in bacterial chromosomes on three aspects as follows–

A. Intra-strand parity in chromosomes

- i. To develop a method for studying intra-strand parity violation in chromosomes.
- ii. To investigate the causes of parity violation in chromosomes with respect to three phenomena
 - (a) GC skew and AT skew in chromosomes.
 - (b) Gene distribution asymmetry in the two complementary strands of the DNA.

(c) Asymmetry in the replication topography.

B. Influence of strand specific mutational bias on codon usage bias

To study the influence of strand specific mutational bias on highly and weakly expressed genes in *Escherichia coli*, the bacterium with strong selected codon usage bias.

(a) Arrangement of genes according to their expression level from Ishihama *et al.* (2008).

(b) Comparison of change in relative synonymous codon usage (CRSCU) between the LeS and the LaS of highly and weakly expressed genes.

C. Strength of selected codon usage bias

- i. To find out strand specific mutational bias in bacterial chromosomes.
- ii. To compare the strand specific mutational bias with the selected codon usage bias 'S' given by Sharp *et al.*, (2005).
- iii. To compare the strand specific mutational bias with UCU(g) (a new measure of selected codon usage bias developed by us) using correspondence analysis and effective number of codons.
- iv. To compare between 'S' and UCU(g).

Acknowledgement

First of all, I express my heartiest gratitude to Prof. Munindra Borah, my supervisor and Dr. Suwendra Kumar Ray, co-supervisor for their guidance and encouragements that I got in my PhD research. Here I would like to acknowledge everyone who contributed directly or indirectly for the fulfilment of this research work. Starting research in Bio-informatics was an accidental event in my life. In 2005, I was looking for a research field involving applications of statistical tools, targeting a field in developmental economics. One day accompanying my friend Dr. Prabin Kalita, I visited Dr. B. Saharia, Controller of Examinations Tezpur University (TU), who was in fact our statistics teacher during my B. Sc. in Darrang College. We were discussing research fields that may be suitable for me for taking up a career in research. In the mean time, Prof. A. K. Buragohain, faculty in the Department of Molecular Biology and Biotechnology (MBBT), (presently Registrar, Tezpur University) visited the Controller of Examinations. He listened to me with patience and explained the scope of research for a statistician in Bio-informatics. Both Prof. Buragohain and Dr. Saharia suggested me to do research work in this inter-disciplinary field. They introduced me to Dr. Suwendra Kumar Ray a young faculty in the Dept. of MBBT, Tezpur University. They suggested to me to apply for enrolment as a part time research fellow in the Dept. of Mathematical Sciences. I approached Prof. M. Borah, Dept. of Mathematical Sciences, Tezpur University, seeking admission as a PhD scholar under the joint guidance of Prof. Borah and Dr. Ray. Finally I was allowed by the DRC of Math. Sciences to undertake the research in this inter-disciplinary field of Biology, Statistics and Computer Science & Information Technology. I started PhD work under the joint guidance from January 2006. The dramatic incidence of meeting Prof. Buragohain has resulted in the form of this thesis. Research in Bio-informatics, as I see it now, was like a journey into a new world contained in a living cell. Though I could not go much deeper into it, only a tip of iceberg that I could touch has given me enormous pleasure of learning.

I am grateful to Prof. M. Borah for providing me research opportunity under his guidance. He was always kind and helpful to me. I learned the basics of computer and research methodology from him. Without any supervisory hindrance he always gave me full autonomy and encouragement to try different methods useful in my research. In the initial stage of learning basics of molecular biology I found a lot of problems. His spirited inspiration helped me to keep patience.

I feel a shortage of words while acknowledging the part of my next supervisor Dr. S. K. Ray. He taught me from the basics of molecular biology, the DNA structure and applications of computational tools. I will remember throughout my life those valuable moments that we shared together exchanging our views. In the initial stage the drawing room of his residence used to be the venue of our week end discussions. I shall never forget the warm hospitality that we were always provided by his wife Mrs. Shindhu Nandini Ray in the course of hours long discussions. From the beginning of this research work I found her passionate towards our exercises for excellence. Dr. Ray was my teacher, my friend and I found him equally eager to learn statistics from me like a student. It was his great venture to develop a platform for Bio-informatics research in this University and I hope with the enthusiastic participation of future scholars he will be able to develop it into a stage of global standard. He remains busy in the department from morning 8 O'clock to evening 8 O'clock devoting himself to the service of students. God may always help him in keeping the enthusiastic teacher within himself. I cannot describe with words the passion he keeps for science. He wants his nation to progress in the field of science and technology and he believes that by strengthening the student-teacher relationship that goal may be achieved. There will remain a little scope in my part to support his dream as soon as I join my job after completion of the thesis work. I hope we will be able to continue our collaboration in future also to do some quality works.

I would like to acknowledge the role of Dr. Prabin Kalita of Darrang College, Tezpur. It was due to his constant inspiration that finally I came out for research work. Dr Kalita helped on many occasions during my PhD. I shall remain ever grateful to him for this rare virtue of friendship.

I am grateful to Dr. B. Saharia, Controller of Examinations, and Prof. A. K. Buragohain, Registrar, TU, as it was the result of their foresighted suggestion that I got an opportunity to learn few things of this rapidly expanding horizon of science and technology.

I shall remember the enormous facilities that I got in the two Departments of Mathematical Sciences, and Molecular Biology and Biotechnology (MBBT) of Tezpur University. In the course of my study I came into contact with all the faculties in the two Departments. I was inspired by all and bestowed with useful suggestions. Most part of my research works were done in the MBBT Department. The internet facility of the University, e-resources available in the Inlibnet site and facilities in the Central Library of the University

were of great help to me. In addition I shall remember the financial support that I was offered by the University to attend the international conference in Hyderabad University and to remit the open access charge for the research article publication in the journal "DNA Research".

I am grateful to the internationally reputed scientists D. R. Forsdyke, M. dos Reis, E. P. C. Rocha, S. K. Kar and other anonymous reviewers for their suggestions and criticisms on our manuscripts. These suggestions gave me direction in the research work. I like to acknowledge the friends of Dr. S. K. Ray working in different labs of the world who helped us by sending valuable research articles which we could not access from Tezpur University. Some intellectual resources in the form of free packages in the web helped me a lot in the computational works. Among these, software packages codonW, clustalW, DNA for windows, PHYLIP, sequinR and trial version of XLSTAT were useful to me. I am grateful to John F Peden for his package codonW kept as a resource in the public domain of the web which can be accessed by anybody. One of the highly cited resources in the publications of the scientists round the globe, codonW serves the purpose of a researcher in this field. The resourceful web pages of the scientists J.R. Lobry, E.P.C. Rocha, D. R. Forsdyke, S. Karlin, P.M. Sharp and many others were among the cites those I used to visit regularly. I am highly indebted to the resources freely available in the websites of DDBJ (DNA data bank of Japan), CMR (Comprehensive microbial resources) and NCBI. The book titled 'Molecular Evolution' by W.H. Li (available in the Central Library, TU) made me benefited giving detailed information about the trends of research and the methodologies used in the studies of molecular evolution.

I am grateful to Prof. B. K. Konwar, Dean of School of Science and Technology; Prof. D. Saikia, Dean of School of Computer Science and Engineering, Prof. N. Dekabarua, Head of the Dept. of Mathematical Sciences and Prof. A. K. Mukherjee, Head of the Dept. of MBBT for their help and support that I got at different times. The seminars towards the end of every Semester in the Department of Math. Sc. were held with sufficient importance and care. These seminars have provided us opportunities of meeting all the scholars from diverse fields of studies ranging Number theory, Fuzzy mathematics, Computer algorithm, Developmental economics, Stochastic processes, Bio-informatics etc. I would always remember Prof. M. Borah, Dr. D. Hazarika, Dr. M. Hazarika, Mr. Bhim Sharma, Mr. Shantanu Dutta and HoD Prof. N. Dekabaruah and other faculty members of the Department for their active participations and scholarly comments on the topics. I have found Ajanta Ba,

Jonali, Chumchum, Bipul, Abhijit, Surabhi and all my juniors to be interested in the topics that I used to present.

I got a lot of support and inputs from my fellow research scholars in the Departments of Computer Sc. and Engineering, Mathematical Sciences and MBBT. Mr. Siddartha S Satapathy from Computer Science helped me by writing 'c' programs necessary for the study. Both of us worked together for the study of "selected codon usage bias" in the Chapter IV. Aditya and Pankaj former MSc students of MBBT helped me in handling and analyzing large sets of electronic data. I also remember Rahul, Pranab, Clara, Sonia and many others from MBBT with whom I interacted on many occasions. I also acknowledge the help and support that I got from research scholars in the department of Mathematical Sciences. I got opportunities to interact with faculty members such as Prof. A. K. Buragohain, Dr. S. Baruah, Dr. M. Mandal from MBBT, TU and Prof. G. Srivastava and Prof. S. C. Kakati from Dibrugarh University in the course of study. Among my colleagues Dr. R. A. Begum, Dr. A. D. Nath and Mr. Rajan Sharma have given me a lot of inspiration and inputs for the work. I am highly grateful to Mr. Debraj Sharma for his help in reviewing the textual and grammatical mistakes in the Preface, Acknowledgement and Abstract part of the thesis.

Next I am going to acknowledge a personality though I did not have any direct contact with him. I see him as a visionary and a mass leader who can influence the people at grass root level. His presence in the Tezpur University has changed its identity. Within a short period of three years he was able to keep this University in the right track to fulfil its mission and vision, goals and objectives. In the last three years, all infrastructure bottlenecks, particularly roads and buildings of the University had experienced a tremendous development. I saw the unexpected rapid development of the University in all fronts in the last couple of years. Prof. M. K. Choudhury, present Vice Chancellor of Tezpur University, is now leading this University with all his virtues and processes of modernization. He was able to provide us a transparent administrative set up, a well organized academic environment where, I believe, research and development will be geared up. The University has been benefited much by his efforts of emphasizing quality at all cost in academic matters. It was due to his endeavour that we got a taste of research in modern environment even in this remotest part of the country.

I would like to acknowledge Dr. J. Hazarika, Principal, Darrang College, Tezpur for his help and support towards my research work. I found him passionate towards the activities

of the teachers for professional excellence. I got a lot of encouragement, support and affection from all my fellow colleagues in Darrang College. The curious eyes of the lovely young students in Darrang College make me always fully involved in their activities and in the activities of my professional excellence.

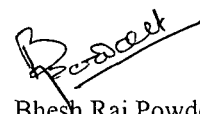
In this hour of submitting my thesis I express my heartiest gratitude to my Statistics teachers Mr. Satyendra Nath Sarmah, Mr. J. D. Goswami and Late Prashanta Kalita. I feel the touch of their blessing hands at all important moments of my life.

I am mostly indebted to my wife Nirmala. It was her dream to see me doing research in a frontier area of science. She put all efforts to relieve me from family matters. She maintained the family and our two children Tushar and Mrinal single handed. It was her constant inspiration that I could finally sit for the thesis. Without her active support it would not have been possible on my part to complete the job.

I feel one important person has not yet been acknowledged is my mother. Although she does not know anything about higher education and research but her face turns satisfied when she sees me busy with computer and research papers. Her smiling face makes me enthusiastic as it used to happen in my childhood.

Last of all I think the 10 km long road connecting Parua in Tezpur town and Tezpur University in Napaam deserves an acknowledgement from me. The road that connects me to the University with its magnificent green coverage is worth mentioning here. The biodiversity lying to its side, the natural beauty it keeps, the life styles of the people living to its side were few things that I enjoyed a lot in the last couple of years. Everyday afternoon that I used to start for Tezpur University from Darrang College on my motorbike, the 25 minutes long journey would refresh me from all sides. It would used to make me a young scholar aged 25 removing all sorts of bitter feelings generated from the activities of day to day life. I always used to get a touch of nature that I lost years back while leaving my village for the job. The small ventures of the poor farmers' that I usually used to see while passing through this road inspired me a lot. I wish the natural beauty of this road may remain for ever in its present form without being converted it to a jungle of concrete. I also remember the warm hospitality of Sarkar, owner of a small restaurant in the Varsity centre. The cup of tea that he used to forward to me with affection was always a refreshing one.

I feel proud of being a PhD student of this University which is rapidly growing into an institution of global standard. Its present exercise to get the status of a true centre for excellence for human resource development makes me applaud with hopes that in near future not only the state of Assam but the entire unprivileged North East of India will be able to experience a rapid social development. Tezpur University will be able to reflect the inherent feelings in the famous song of Jyoti Prasad Agarwala where it is scripted as the voyage of yours and mine towards light (*Tore More Alokare Yatra – Jyoti Prasad Agarwala*).



Bhes Raj Powdel

12.08.2010

Abstract

Chapter I is the review literature which covers three aspects of DNA compositional studies such as Chargaff's 2nd parity rule (PR2), strand specific mutational bias and codon usage bias. In the review of "Chargaff's 2nd parity" we have discussed methodologies used to study PR2 in different genomes and the explanations for such a universal phenomenon in genomes. The review of "strand specific mutational bias" includes the pioneer work of Lobry and Sueoka since 1995 and different mechanisms responsible for the existence of strand specific mutational bias in genomes. In the review of "codon usage bias" different methodologies used to study codon usage bias since 1981 (Ikemura) have been discussed.

Chapter II describes a new methodology named intra-strand frequency distribution parity to study Chargaff's 2nd parity in chromosomes. The important finding in this chapter is that parity violation is commonly observed in bacterial chromosomes. Violation of parity in chromosomes can be attributed to multiple factors operating at different levels.

Chapter III describes the influence of strand specific mutational bias on codon usage of weakly expressed genes in *Escherichia coli*. This work is in support of the selection-mutation-drift theory by taking the information from proteome data.

Chapter IV describes a new approach to study codon usage bias in genes. The approach is named as unevenness of codon usage (UCU). Using correspondence analysis it has been shown that the approach gives information about selected codon usage bias in genes in bacteria. The results were also compared with the effective number of codons (ENC).

Contents

1. Literature review	18
1.1. Abstract	18
1.2. Chargaff's rules of nucleotide composition in DNA molecules	18
Figure 1.1: DNA double helix model	19
1.2.1. Methodologies to study of Chargaff's 2 nd parity in chromosomes	20
1.2.1.1. Studies using regression analysis on sample DNA fragments	21
1.2.1.2. Whole genome composition studies using t-test	21
1.2.1.3. Symmetry studies using Markov chain	22
1.2.1.4. 2D DNA walk method	23
1.2.2. Explanations for the observation of PR2 in chromosomes	23
1.2.2.1. Stem-loop hypothesis	23
1.2.2.2. Inversion and inverted transposition hypothesis	24
1.2.2.3. Parity rule 2 (PR2) under no strand bias condition	25
Figure 1.2: Rates of base substitution	26
1.3. Replication and composition of coding sequences	27
1.3.1. Strand specific mutational bias	27
Figure 1.3: Cumulative ATS and GCS in <i>B. subtilis</i> (top) <i>E. coli</i> (bottom)	29
1.3.2. Causes for strand specific mutational bias in genomes	30
1.3.2.1. Gene distribution asymmetry and the role of DNA polymerase	30
1.3.2.2. Replication gradient	31
1.4. Genetic code and synonymous codon bias	32
1.4.1. Earlier studies on codon usage bias	32
1.4.2. Measures of codon usage	33
1.4.2.1. Fop	34
1.4.2.2. P2	34
1.4.2.3. RSCU	34
1.4.2.4. CAI	35
1.4.2.5. ENc	35
1.4.2.6. Shannon information based codon bias	36
1.4.2.7. ICDI	37
1.4.2.8. tAI _g	37
1.4.2.9. Correspondence Analysis	38
1.4.3. Underlying hypotheses for codon usage bias	39
1.4.3.1. Hypothesis based on natural selection	39
1.4.3.2. Hypothesis based on mutation	39
1.4.4. The Selection-Mutation-Drift theory and Population Genetics model	40
1.5. Discussion	42

2. A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (Chargaff's 2nd parity rule).....	44
2.1. Abstract	44
2.2. Introduction.....	44
2.3. Materials and Methods.....	47
2.3.1. Frequency distribution calculation	47
2.3.2. Proportionate distribution of forward encoded and reverse encoded sequences in a DNA strand.....	48
2.3.3. Calculation of whole genome AT skew and GC skew	48
2.3.4. Identification of leading and lagging strand region.....	48
2.3.5. Relative proportion of coding sequence distribution.....	49
2.4. Results	49
2.4.1. Intra-strand frequency distribution parity in chromosomes of bacteria.....	49
Table 2.1: Result of Kolmogorav-Smirnov (KS) test for significance between the frequency distribution of complementary nucleotides.	50
Table 2.2: Summary of the frequency distribution parity test.....	58
Figure 2.1 (a-e): Frequency distribution of nucleotides in chromosomes.....	59
2.4.2. ISFDP weakly correlates with Chargaff's 2 nd parity	66
2.4.3. The chromosomes with asymmetric replication topography are more prone to ISFDP violation in bacteria	67
2.4.4. Composition of forward encoded and reverse encoded sequences within DNA strands might influence the parity.....	68
Figure 2.2: Schematic representation of coding sequence arrangement studied.....	70
Figure 2.3 (a, b) : Frequency distribution study of nucleotides in coding sequences ...	71
Figure 2.4: Relative disproportionate composition of ORFs between Ws and Cs in Chromosomes.....	72
2.4.5. Intra-strand frequency distribution parity between complementary oligonucleotides in chromosomes	73
Figure 2.5(a, b): Frequency distribution of dinucleotides	74
Figure. 2.6(a,b): Frequency distribution of trinucleotides in Escherichia coli chromosome	75
2.5. Discussion	76
3. Strand-specific mutational bias influences codon usage of weakly expressed genes in Escherichia coli.....	81
3.1. Abstract	81
3.2. Introduction.....	81
3.3. Materials and Methods.....	84
3.3.1. Separation of highly expressed genes and weakly expressed genes in LeS and LaS ..	84
3.3.2. Codon usage study between LeS and LaS.....	85
3.3.3. ATS ₃ and GCS ₃ between LeS and LaS	86

3.3.4. Estimation of strand-specific mutational bias.....	87
3.3.5. Transfer RNA gene ratio.....	87
3.4. Results	87
3.4.1. Study of synonymous codon usage bias in ascending order of gene expression.....	87
Table 3.1: Correlation in the expression rank order and the CRSCU	88
Table 3.2: Correlation in the expression of rank order (revised group) and the CRSCU.....	89
3.4.2. Strand specific mutational bias between the strands is higher in case of weakly expressed genes than highly expressed genes.....	90
Table 3.3: List of highly expressed genes and weakly expressed genes with their strand location of <i>E. coli</i> MG1655 chromosome analyzed in this study.....	91
Table 3.4: Strand specific mutational bias in codon third position of HEG and WEG and in IR.....	98
3.4.3. ATS_3 and GCS_3 between LeS and LaS.....	98
3.4.4. Higher CRSCU in case of weakly expressed genes than highly expressed genes	99
Table 3.5: Strand specific $ATS_3 = [A_3/(A_3+T_3)]$ and $GCS_3 = [G_3/(G_3+C_3)]$ in highly expressed genes (HEG) and weakly expressed genes (WEG)	100
Table 3.6: Change in relative synonymous codon usage (CRSCU) between the strands.....	101
3.4.5. Higher SCF in case of weakly expressed genes than highly expressed genes	101
Table 3.7: Synonymous codon frequency (Experimental procedures) of family box codons in highly expressed genes and weakly expressed genes	102
3.5. Discussion	103
4. Selected codon usage bias in bacterial chromosomes	107
4.1. Abstract	107
4.2. Introduction	107
4.3. Materials and Methods.....	109
4.3.1. Calculations for UCU(g).....	109
4.3.2. Calculation of strand specific mutational bias in the intergenic regions (mut_ir) as well as at the 3 rd position of codons (mut_3).....	110
4.4. Results	111
4.4.1. Uneven codon usage measure of genes [UCU(g)] in <i>E. coli</i> as well as in <i>S. cerevisiae</i>	111
Table 4.1: Correlation between UCU(g) and CAI in <i>E. coli</i> and <i>S. cerevisiae</i> (yeast).....	113
Table 4.2: Correlation of UCU(g) and CAI with gene expression in <i>E. coli</i> and <i>S. cerevisiae</i> (yeast).....	114
Figure 4.1: A two panel figure presenting scatter plots of UCU(g) vs. CAI.....	114
Figure 4.2: A two panel figure presenting scatter plot of UCU(g) vs. gene expression.....	115
4.4.2 UCU(g) correlates with primary axis of correspondence analysis as well as with effective number of codons (N_c) in several bacteria	115
Table 4.3: Correlation analysis of UCU(g) in 76 bacterial genomes	117

<i>4.5. Discussion</i>	121
5. Conclusion	125
6. References:	127

List of Tables and Figures

Figure 1.1: DNA double helix model.....	19
Figure 1.2: Rates of base substitution	26
Figure 1.3: Cumulative ATS and GCS in <i>B. subtilis</i> (top) <i>E. coli</i> (bottom).....	29
Table 2.1: Result of Kolmogorav-Smirnov (KS) test for significance between the frequency distribution of complementary nucleotides.	50
Table 2.2: Summary of the frequency distribution parity test	58
Figure 2.1 (a-e): Frequency distribution of nucleotides in chromosomes	59
Figure 2.2: Schematic representation of coding sequence arrangement studied	70
Figure 2.3 (a, b) : Frequency distribution study of nucleotides in coding sequences	71
Figure 2.4: Relative disproportionate composition of ORFs between Ws and Cs in Chromosomes	72
Figure 2.5(a, b): Frequency distribution of dinucleotides.....	74
Figure 2.6(a,b): Frequency distribution of trinucleotides in <i>Escherichia coli</i> chromosome.....	75
Table 3.1: Correlation in the expression rank order and the CRSCU.....	88
Table 3.2: Correlation in the expression of rank order (revised group) and the CRSCU.....	89
Table 3.3: List of highly expressed genes and weakly expressed genes with their strand location of <i>E. coli</i> MG1655 chromosome analyzed in this study.....	91
Table 3.4: Strand specific mutational bias in codon third position of HEG and WEG and in IR	98
Table 3.5: Strand specific $ATS_3 = [A_3/(A_3+T_3)]$ and $GCS_3 = [G_3/(G_3+C_3)]$ in highly expressed genes (HEG) and weakly expressed genes (WEG).....	100
Table 3.6: Change in relative synonymous codon usage (CRSCU) between the strands.....	101
Table 3.7: Synonymous codon frequency (Experimental procedures) of family box codons in highly expressed genes and weakly expressed genes	102
Table 4.1: Correlation between UCU(<i>g</i>) and CAI in <i>E. coli</i> and <i>S. cerevisiae</i> (yeast).....	113
Table 4.2: Correlation of UCU(<i>g</i>) and CAI with gene expression in <i>E. coli</i> and <i>S. cerevisiae</i> (yeast)	114
Figure 4.1: A two panel figure presenting scatter plots of UCU(<i>g</i>) vs. CAI	114
Figure 4.2: A two panel figure presenting scatter plot of UCU(<i>g</i>) vs. gene expression	115
Table 4.3: Correlation analysis of UCU(<i>g</i>) in 76 bacterial genomes.....	117

List of Abbreviations

CUB – Codon usage bias	GCS – GC skew
ISP – Intra-strand Parity	HEG – Highly expressed genes
K – G & T nucleotides (stands for keto)	ISFDP – Intra-strand frequency distribution parity
LaS – Lagging strand	PR2 – Parity rule 2
LeS – Leading strand	SCF – Synonymous codon frequency
M – A & C nucleotides (stands for amino)	ssDNA – Single stranded DNA
mut_c3 – Mutational bias at the 3 rd position of codons	dsDNA – Double stranded DNA
mut-ir – Mutational bias in the intergenic region	SSMB – Strand specific mutational bias
ATS – AT skew	UCU – Unevenness of codon usage
CRSCU – Change in relative synonymous codon usage	WEG – Weakly expressed genes

CHAPTER I

1. Literature review

1.1. Abstract

Forsdyke and Mortimer (2000) have reviewed elegantly the impact of E. Chargaff's discoveries in the genomic era. The discovery of the intra-strand parity, $A \approx T$ and $G \approx C$ within individual DNA strands in bacterial chromosomes in 1968, which is now found to be true in genomes of viruses, organelles, bacteria, archaea and eukaryotes, is still an interesting problem for research. An attempt to explain intra-strand parity in genomes led to the discovery of strand specific mutational bias in bacterial chromosomes, which was discovered in 1996 by Lobry. Though strand specific mutational bias is observed in most of the bacterial chromosomes analysed till date, its magnitude varies among the genomes. In fact, our understanding regarding the mechanism of strand specific mutational bias is incomplete. The impact of strand specific mutational bias is so high in certain bacteria such as *Borrelia burgdorferi*, *Xyllela fastidiosa*, codon usage bias in a gene is determined by its mode of replication rather than its expression unlike the case in *Escherichia coli*. This observation opened up another area of research called selected codon usage bias. The strength of selected codon usage bias has been found to be variable among bacteria. The above three issues, Chargaff's 2nd parity, strand specific mutational bias and selected codon usage are still unsolved problems in front of scientists today and are important problems to work with for a researcher interested in genome composition. The objectives of the thesis are based on these three topics. The major developments in the three issues such as Chargaff's 2nd parity rule, strand specific mutational bias and codon usage bias in bacteria have been described in this chapter.

1.2. Chargaff's rules of nucleotide composition in DNA molecules

Chargaff's 1st parity rule (Chargaff, 1950, 1951) based on nucleotide composition of double stranded DNA states that the complementary nucleotides have the same abundance values i.e. $f_A = f_T$ and $f_G = f_C$ where $f_A + f_T + f_G + f_C = 1$ (Forsdyke and Mortimer, 2000). This is explained by the DNA double helix model (Fig. 1.1) in which A pairs only with T, and G pairs only with C (Watson and Crick, 1953). The compositional similarity between complementary nucleotides in a DNA duplex does not give information about the

compositional relationship between complementary nucleotides within individual strands in the DNA molecule.

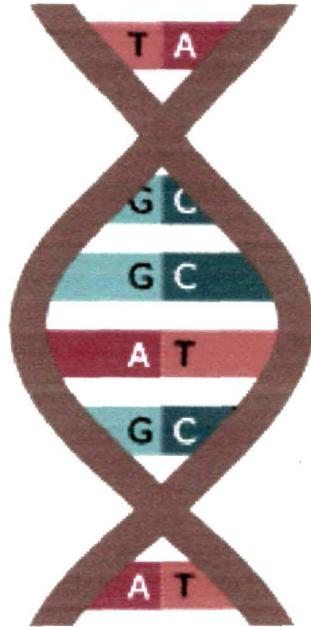


Figure 1.1: DNA double helix model

The chemical composition of individual DNA strands was also reported in 1968 from Chargaff's laboratory for *Bacillus subtilis* chromosome and extended later to six more bacterial species (Rudner *et al.*, 1968; Rudner *et al.*, 1969). Chargaff and his colleagues observed the similarity between the abundance values of complementary nucleotides ($f_A \approx f_T$, $f_G \approx f_C$) within individual DNA strands of bacterial chromosomes, which was very surprising for them. In the post genomic era, the compositional similarity between complementary nucleotides is observed in chromosomes of bacteria, archaea and eukaryotes, which is now known as Chargaff's 2nd parity or intra-strand parity (ISP) (Forsdyke and Mortimer, 2000) or parity rule 2 (PR2). In the following sections these terms have been used alternatively for Chargaff's 2nd parity.

1.2.1. Methodologies to study of Chargaff's 2nd parity in chromosomes

In 1970s, with the development of DNA sequencing methodologies, data-base of DNA sequences of bacteriophage and some of the organelles came to exist. The data base attracted scientists to undertake compositional studies of the DNA molecule. Such studies were primarily related to sequence alignment, gene finding, DNA compositional studies, codon usage studies etc. (Grantham *et al.*, 1980; Ikemura, 1981; Gouy and Gautier, 1982; Bennetzen and Hall, 1982; Bernardi *et al.* 1985; Bernardi and Bernardi, 1986; Sharp and Li, 1986a, 1986b; Bulmer 1987, 1991). Although intra-strand parity in chromosome was discovered in the year 1968 (Rudner *et al.* 1968), extensive research in this field are found in the 1990s.

The simple approach to study PR2 is to find the total abundance of individual nucleotides within a DNA strand and calculate AT skew (ATS) = $(A-T)/(A+T)$ as well as GC skew (GCS) = $(G-C)/(G+C)$. In an ideal case of PR2 both GCS and ATS tend to zero because of the similar abundance values of complementary nucleotides. But Szybalski and his colleagues (1966) had already described that purine richness predominated in the coding strand and pyrimidine richness predominated in the non-coding strand. In support of this, Smithies *et al* (1981) reported strand compositional asymmetries in 11,376 nucleotides of sequenced DNA from the human fetal globin gene region. The authors divided the region into 113 segments, each of approximately 100 nucleotides, and looked at the compositional asymmetries with each division. They observed significant local variation in the strand asymmetries along the length of the sequences, irrespective of whether or not strand asymmetries are accepted in the sequence as a whole. Though the study here was done in human gene, the findings were in support of the study done earlier on bacterial genomes by Szybalski *et al* (1966). As more gene sequences were available, it became clear that local violation of PR2 is a rule rather deviation in genomes. As genome sequences were available, PR2 was analyzed in whole genomes. Shioiri and Takahata (2001) analysed 152 complete mtDNA sequences, 36 complete prokaryote chromosomes, and several long contigs for human and *Arabidopsis thaliana* chromosomes using ATS and GCS measures. This study had reported that in most organisms, excluding some invertebrates and plants, ATS and GCS over the whole mitochondrial genomes often deviates significantly from zero, and the

absolute ATS and GCS values differ from each other. All 36 prokaryote chromosomes showed that ATS and GCS in the entire region are almost zero.

1.2.1.1. Studies using regression analysis on sample DNA fragments

A generalized presentation on the Chargaff's 2nd parity was given by Prabhu (1993). He studied all the available sequences in the GenBank which are 50000 base pair or more. Using linear regression charts, the observed parity in the frequencies of the complementary bases was shown. Observed parity in the complementary oligonucleotides up to 6th order was presented with Pearson's correlation coefficients and linear regression coefficients. The complementary frequencies in all the cases were similar and he got correlation coefficients approaching unity and all the regression plots were passing through the origin having unit regression coefficients. But he was silent in explaining such enigmatic property of the DNA sequences.

Mitchell and Bridge (2006) analyzed 231 bacterial chromosomes, 1495 viral genomes, 835 organellar genomes, 20 archaeal genome, 164 sequences from 15 eukaryotes to test Chargaff 2nd parity using regression analysis. They reported that PR2 is true for all double stranded DNA with the exception of organelle genomes. In addition violation of PR2 was observed in single stranded viral genomes. PR2 study on organelle genomes was further done by Nikolaou and Almirantis (2006). According to them most of mitochondrial genomes exhibited PR2 violation whereas chloroplast genomes exhibited parity. They studied the violation of PR2 in organelle genomes using measure based on GCS and ATS which is given by $d(\text{PR2}) = \sqrt{(\text{ATskew}^2 + \text{GCskew}^2)}$ as a measure of deviation from parity (Nikolaou and Almirantis, 2006).

1.2.1.2. Whole genome composition studies using t-test

Qi and Cuticchia (2001) studied PR2 in 26 prokaryotic chromosomes and 8 eukaryotic chromosomes (Qi and Cuticchia, 2001). Like Prabhu, they also used linear regression plots and correlation analysis to study similarity of complementary bases and reverse complements of di and tri nucleotides in a single strand of chromosome. To test the significance of the paired similarity they used t test. Baisnée *et al.*, (2002) has criticized this approach of using paired t-test for testing the significance of symmetry. The null hypothesis

of no difference in the counts of Nmers and their reverse complements makes it irrelevant for asymmetric distribution. Biological symmetry is not perfect enough to fulfil the required assumptions for a statistical test. Rather they have advocated the analytical approximations and simulations to estimate the symmetry of the distributions (Baisnée *et al.*, 2002).

1.2.1.3. Symmetry studies using Markov chain

Baisnée *et al.*, (2002) made an attempt to measure the symmetry (PR2) from mononucleotide to ninth order oligonucleotide level across a wide set of genomes ranging from ssDNA and dsDNA of viruses, bacteria, archaea, mitochondria and eukaryotes. They put an effort to investigate interdependence of the parity in higher order and lower order oligonucleotides. The prime methodology used in their work was linear regression plots of the $4^N N^{\text{th}}$ ($N= 1, 2 \dots 9$) order oligonucleotide (Nmer) frequencies along a given DNA strand against the similar frequencies in the complementary strand. Such plots were in general symmetric with respect to the main diagonal line showing parity between an oligonucleotide and its reverse complement. To measure symmetry quantitatively two indices namely S^1 and S^C were used where S^1 was defined as complement to unity weighted average of the absolute values of the skews of all Nmer reverse complement frequencies (f_i and f_i^1) along a DNA strand where the weights were taken as $(f_i + f_i^1) / \sum_i (f_i + f_i^1)$. S^C was defined as the Pearson's correlation coefficients between f and f^1 . S^1 ranges from 0 to 1, S^C ranges from -1 to 1. Statistical Markov models were used to analyse the origin of the phenomenon of symmetry. The sole objective in using statistical Markov models was to see whether symmetry in higher order is obtained as a consequence of the symmetry at the lower order or vice-versa. Analysing strand symmetry across taxa the authors have put forward the view that symmetry increases in a consistent manner with sequence length both across and within genomes. Distribution of symmetry levels across length is having some similarity which has led the authors to accept that strand symmetry in polynucleotide molecule is an emerging property under evolutionary pressures. Moreover the actual symmetry levels in biological sequences were found to be lower and more variable than those obtained using statistical models. The phenomenon of strand symmetry has been considered in the article as an outcome of the compound effects of a wide spectrum of mechanism operating at multiple

orders that tends to shape the two complementary strands functionally similar and doesn't represent a direct constraint or add a selective advantage. The authors also point out the biases in the gene distribution between strands that may lead to the first order asymmetry.

1.2.1.4. 2D DNA walk method

In the 2D (2-dimensional) DNA walk method a DNA sequence is mapped into the square lattice on the plane with GC and AT axes, where the origin (0,0) coincides with the first nucleotide in DNA sequence (Poptsova *et al*, 2009). 2D DNA walk is a method of DNA sequence representation on a plane whereby a trajectory is drawn, nucleotide after nucleotide, in four directions: G-up, C-down, T-left, A-right. Chromosomes show composition complexity change from symmetrical half-turn in bacteria to pseudo-random trajectories in archaea, fungi and humans. Transformation of gene order and strand position returns most of the analyzed chromosomes to a symmetrical bacterial-like state with one transition point. Results in this study shed light on the Chargaff's 2nd parity rule that was previously applied to DNA sequence containing both genes and intergenic regions. Here it is demonstrated that this rule holds true for DNA sequences made up solely of genes and is strongly correlated with the equal number of genes on strands. Besides, this study shows that the absence or presence of nucleotide skews in chromosomes can be explained by the location of genes on strands, and that the majority of the investigated genes (coding sequences) are G and A rich.

1.2.2. Explanations for the observation of PR2 in chromosomes

1.2.2.1. Stem-loop hypothesis

The stem-loop hypothesis is commonly known as Nussinov-Forsdyke hypothesis. The main point of this hypothesis is that there is a genome wide selection for formation of DNA secondary structure (DNA stem-loop regions) which is advantageous to the cell for processes like recombination (Lobachev *et al*. 1998). Formation of DNA secondary structures is the main selection force for the observation of PR2 in genomes. Early works of Nussinov (Nussinov, 1982; Hinds and Blake, 1984, 1985) are in support of the DNA structure model. Thus, a sequence containing an inverted repeat (e.g. NNNATGNNNCATNNN) has palindrome-like characteristics with the potential to fold back on itself forming a stem-loop, hairpin-like, structure. Wherever this structure appears, then ATG = CAT, suggesting why

the frequency of ATG is equal to the frequency of its inverse complement. The literature shows that, especially when negatively supercoiled, duplex DNA will adopt stem-loop (sometimes cruciform) configurations and correlating with their high content of inverted repeats, DNA molecules from biological sources show a general potential to extrude such higher ordered structures. New technologies have allowed direct visualizations of this (Woodside *et al.*, 2006). Irrespective of the selective forces that led to such structures, their existence provides some explanation for Chargaff's 2nd parity. Forsdyke and Mortimer (2000) concluded that organisms that had accepted point mutations which increased the probability of stem-loop formation (both in protein-coding and in non-protein-coding DNA), had usually had an evolutionary advantage over organisms which had not accepted such mutations.

1.2.2.2. Inversion and inverted transposition hypothesis

There are two independent publications suggesting genome wide inversions are responsible for the establishment of parity in chromosomes (Albrecht-Buehler, 2006; Okamura *et al.*, 2007). Albrecht-Buehler (2006) has viewed Chargaff's 2nd parity as an outcome of presence of million copies of interspersed repetitive elements in the genome and genomes have no selective advantage in complying with PR2 (Albrecht-Buehler, 2006). According to Albrecht-Buehler (2006), PR2 is not an outcome of the statistical regularity expected in case of long natural sequences. The prime methodology used in his work was count statistics of the triplets and their reverse complement in the same strand under the assumption that the two strands are homogeneous in nature. He was with the opinion that complying with PR2 for mononucleotides doesn't necessarily imply complying with oligonucleotides although the reverse may be true. Correlation plots were used to quantify the degree of compliance of the genomes with PR2. Analysing more than 500 genome segments of length 8Mb or smaller he found only a subset of mitochondrial genomes violating PR2. It was assumed that all genomes initially violated the PR2 because they contained arbitrary number of single nucleotide. Only the subsequent evolution rendered them comply with PR2 in case of mono and oligonucleotides. Mechanism responsible for this was assumed to be inversion and inverted transposition. Insertion of chromosome sections in reverse order in their original location is called inversion or inserting somewhere else is known as inverted transposition. These activities inside a chromosome are far to swap strands. A particular

section which was a part of Watson strand has to be inserted in to Crick strand and vice-versa. These actions gradually equalize the complementary nucleotides in one strand. The process is self stabilizing and once the genome complies completely with PR2 this property is maintained forever. Thus the author was with the opinion that compliance of the genomes with PR2 is an inevitable and asymptotic in the course of evolution.

Okamura *et al.* (2007) has viewed the second parity in the chromosomes as a result of genome wide occurrences of repeated inversions. With the help of a mathematical limiting model, theoretically they have shown that after 'n' repetition the frequencies of A and T may be shown as

$$\lim_{n \rightarrow \infty} A_n = \frac{1}{2}(A_0 + T_0) \text{ and } \lim_{n \rightarrow \infty} T_n = \frac{1}{2}(A_0 + T_0)$$

Where A_0 and T_0 are the initial frequencies of A and T respectively.

1.2.2.3. Parity rule 2 (PR2) under no strand bias condition

Sueoka (1995) studied intra-strand parity in synonymous third codon positions, the selectively neutral sites of a genome. Introducing two types of parities namely PR1 and PR2, his objective in the study was to analyse the relative role of directional mutation pressure and selective codon usage bias on the violation of PR2 in the coding region. PR1 was concerned with the base substitution rates in individual DNA strand while PR2 was concerned with base composition in individual DNA strand. Intra-strand substitution rates determine the relative frequencies of each nucleotide A, C, G and T in a single strand. In a strand bias situation, there are twelve different possible mutation rates between four bases of nucleotides which reduce to six under no strand bias condition (PR1; Fig. 1.2). Up to the year 1994 no complete genome sequences of bacteria were available in the GENBANK and the PR2 studies made by Prabhu (1993) were based on the DNA sample of 50000 or more bases where the effects of local asymmetries and non randomness cannot be nullified. Sueoka justified the study of violation of PR2 by taking coding regions from different locations and studying their asymmetry in the third codon position. He plotted the ATS and GCS at the third codon position for the eight family boxes in the genes of the organisms comprising eukaryotes to prokaryotes and came with the conclusion that "violation of PR2 is the rule rather than

exception, and the violation pattern is unique for each of the eight amino acids and distinctly different between two organisms". He was with the view that the correlation between tRNA abundance and the synonymous codon frequency is a general cause for base composition asymmetry leading to PR2 violation in sense strand.

In an accompanying paper using model of DNA evolution Lobry (1995) put forward the view that intra-strand equimolarity between A and T and between G and C is a general asymptotic property of the model based on the assumption of no strand bias (Lobry, 1995). Sueoka (1999) presented another finding which shows that genes in different GC content groups are similar with respect to PR2 violation. This was an important finding because genes of higher eukaryotes are located in the isochores of heterogeneous GC. This indicates that directional mutation pressure and translational selection led violation from PR2 are uncorrelated.

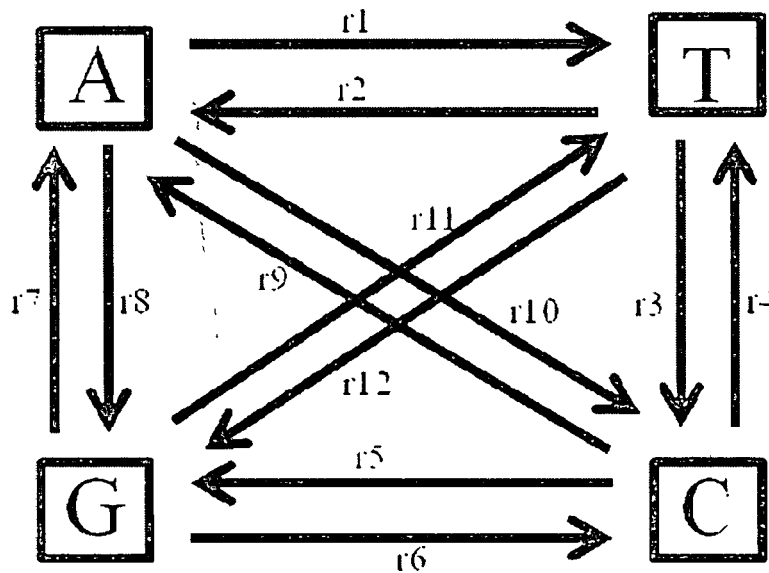


Figure 1.2: Rates of base substitution

The 12 substitution rates of bases in DNA: r_3 , r_4 , r_7 and r_8 are transitions (4 ways). The others (r_1 , r_2 , r_5 , r_6 , $r_9 - r_{12}$ are transversions (8 ways). The 12 substitution rates determine

the nucleotide composition within individual DNA strands. The 12 substitution rates can be converted to six substitution rates considering the complementary base pairing rule (Parity rule I; PR1). Using the substitution rates under no strand bias condition i.e. $A \rightarrow T (r1) = T \rightarrow A (r2)$. It can be easily deduced that $A = T$ as well as $G = C$ even within individual strands in a DNA molecule and this relationship is called as intra-strand parity or parity rule II (PR2). This description is taken from Sueoka (1995).

1.3. Replication and composition of coding sequences

Local violation of PR2 occurs due to replication, transcription and translation in genomes. PR2 violation is observed in small DNA regions though the entire genome exhibits parity. This is due to the cancellation effects of the parity violations in both directions. Mitchell and Bridge (2006), as well as Nikolaou and Almirantis (2006) have described the higher distribution of coding sequences in one of the strands in organelle genomes is an important reason for the violation of parity in these genomes. This is also true for the single stranded bacteriophages for the violation of parity. In organelle genomes the replication process is different which causes the biased distribution of gene sequences between the strands that result into the violation of parity. Baisnée *et al.* (2002) have described that violation of parity involves multiple reasons and no single reason is sufficient to describe the violation of PR2 in bacterial chromosomes.

1.3.1. Strand specific mutational bias

Under no strand bias between the LeS and LaS with respect to mutation and selection, the composition of complementary nucleotides within a DNA strand will remain similar, which is known as PR2 or intra-strand parity (Sueoka, 1995, Lobry 1995). However, the asymmetry during DNA replication has been shown to affect differentially mutation/nucleotide-substitution rates between the strands. Wu and Maeda (1987) were the first to report the inequality in mutation rates of the two strands of DNA. Their conclusion is based on the aligning of homologous sequences in a region of the β -globin complex of primates and estimating the substitution matrix and comparing the frequencies of complementary changes. However, the origin and terminus of DNA replication were not defined in their studies for which the observation was applicable (Frank and Lobry, 1999). Due to single origin of replication, bacterial chromosomes are more suitable for comparing

the substitution patterns between the two strands in a DNA molecule. Lobry used $GCS=(C-G)/(C+G)$ and $ATS=(A-T)/(A+T)$ over sliding windows along a DNA sequence to prove the existence of GC and AT skews in the genomes of *Haemophilus influenza* and in parts of *Escherichia coli* and *Bacillus subtilis*. In these bacteria the skews switch sign at the origin and terminus of replication. LeS is observed generally richer in G than C and in T than A, *vice versa* for the LaS. Grigoriev (1998) presented the genome in the form of a cumulative skewed picture which resulted into a 'v' shaped or inverted 'v' shaped structure as shown for *Escherichia coli* and *Bacillus subtilis* (Fig. 1.3). The strand specific mutational bias is found in bacterial genomes as well as in viral genomes (McLean *et al*, 1998; Mřazek and Karlin, 1998; Kano-Sueoka *et al.*, 1999). There are several experimental studies that demonstrate differential mutation rates between the two strands (Frank and Lobry, 1999). A study by Fijalkowska *et al.* (1998) in an *E. coli* chromosome that involves the measurement of *lac* reversion frequency by base substitution for the two orientations reported that the lagging strand is more accurate than leading strand. Mismatch and proofreading deficient strains were used to detect intrinsic error rates between the strands.

The discovery of strand specific mutational bias enabled the scientists to predict the potential origin and termination sites of replication with the help of the skew switches in bacterial chromosomes. Based on the skew pattern the origin of replication was predicted in the chromosome of *Borrelia burgdorferi* which was later proved to be correct by experiments (Frank and Lobry, 1999). A computer based program named Oriloc was made initially to predict origin and terminus of a chromosome based on the skews (Necşulea and Lobry, 2007). Now other programs such as CG-software, GrapfDNA, Z-curve, Oligonucleotide skew method and Ori-finder, DoriC database are also available to predict origin and terminus of replication in bacterial chromosomes (Sernova and Gelfand, 2008).

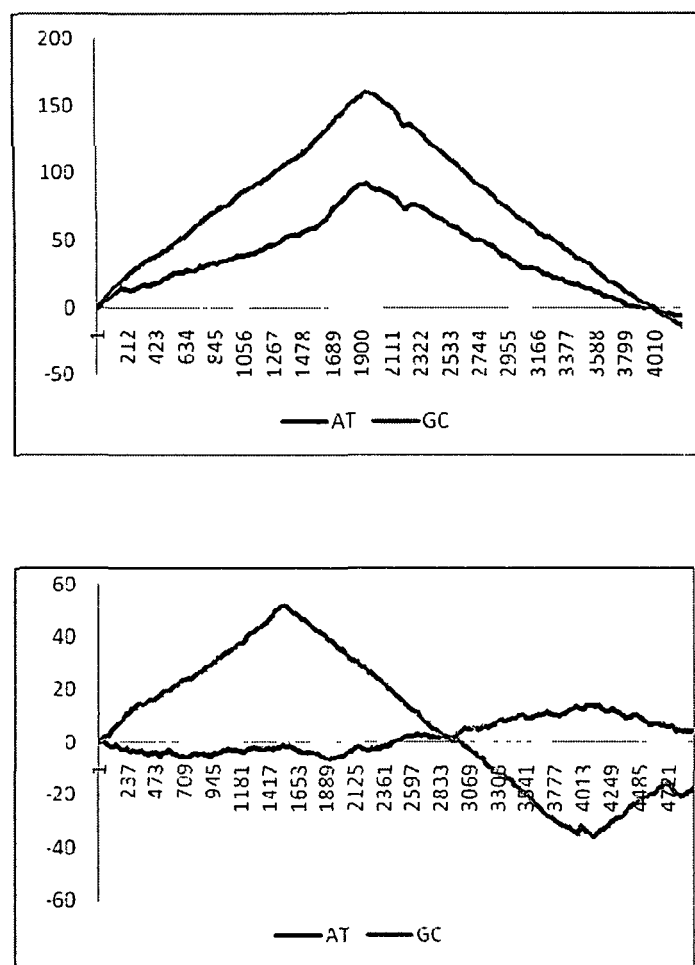


Figure 1.3: Cumulative ATS and GCS in *B. subtilis* (top) *E. coli* (bottom)

X-axis represents the chromosomal coordinate in kb, *Y*-axis represents cumulated skews (ATS, GCS) in every 1kb window starting from the first position of the chromosome. In *B. subtilis* both ATS and GCS exhibits similar patterns whereas in *E. coli* the patterns are opposite. Cumulated GCS is generally found increasing along *LeS*. The GCS changes its

polarity at ori and ter of chromosome replication in both organisms. ATS skews are not found always with the similar pattern in all chromosomes.

1.3.2. Causes for strand specific mutational bias in genomes

Strand specific mutational bias is caused by both replication and transcription. Cytosine deamination in single stranded DNA is 100 times more frequent than double stranded DNA (Francino and Ochman, 1997). Higher cytosine deamination of exposed DNA as single stranded during replication and transcription are the main causes for the strand specific mutational bias (Gautier, 2000; Francino and Ochman, 1997, 2001; Green *et al.*, 2003). In addition the transcription coupled repair, which acts only on the template DNA, also contributes to the strand specific mutational bias (Francino *et al.*, 1996).

Though the ATS sign is usually reverse to that of the GCS sign in a strand, the observation is not universal. In Firmicutes (gram-positive bacteria with low GC%), ATS and GCS signs are same with respect to a strand (Freeman *et al.*, 1998; Hu *et al.*, 2007). Whatever the sign of the skews may be, their effect is very prominent. Moving along a chromosome sequence in a window, the skew switches the symbol (e.g. +ve to -ve) at terminus as well as at the origin of replication. Using this computational approach the origin of replication was predicted for *Borrelia burgdorferi*, which later turned to be true. The skew is presented in a better way by plotting the cumulative addition value against the coordinates starting from the beginning of the sequence to the end of it (Griegoreiv, 1998). The PR2 plot derived by Lobry and Sueoka (2002) is a pictorial presentation of the asymmetries in the composition of coding regions in LeS and LaS. These plots along with the measures of B_I and B_{II} (described in Chapter III) help in separating the biases in base composition due to (i) replication associated mutational bias and (ii) transcription/translational associated biases (Lobry and Sueoka, 2002).

1.3.2.1. Gene distribution asymmetry and the role of DNA polymerase

The LeS and LaS are also asymmetric in terms of gene distributions. Genes are preferably located in the LeS than the LaS to reduce collision between the machineries of replication and transcription (Rocha, 2004). Degrees of the asymmetry between the strands vary among chromosomes and are dependent upon the composition of DNA polymerase III.

The DNA polymerase in *E. coli* (*E. coli* type) consists of two identical units of DnaE, which are involved in synthesizing leading and lagging strands. Whereas in case of *B. subtilis* (*B. subtilis* type) it is made up of two different units: DnaE enzyme involves in synthesizing LaS and PolC involves in synthesizing the LeS. DnaE lacks an error repairing system while PolC possesses it. Usually a bacterium possesses either *E. coli* type or *B. subtilis* type DNA polymerase. Analysis of different bacterial chromosomes has revealed that in organisms with *B. subtilis* type polymerase, a higher asymmetry of gene distribution is observed between the strands than the organisms with *E. coli* type polymerase. For example, the gene distribution between the LeS and LaS in *B. subtilis* is 74% and 26% respectively, whereas the same in the case of *E. coli* is 55% and 45%. The basic difference in the replication machinery causes different degree of asymmetry between the strands in chromosomes. The other asymmetry between LeS and LaS is the distribution of the type of genes between the strands. There are two views regarding this asymmetry. First, 'gene expressivity' according to which highly expressed genes are preferentially located in LeS than LaS. Second, 'gene essentiality' according to which essential genes are preferentially located in LeS than LaS. Gene essentiality holds the opinion that expression of a gene will not be affected significantly in either of the strands. Rather abortive transcript and dominant -ve effect of faulty proteins are the major reasons for the gene distribution asymmetry. The reason for gene distribution asymmetry between the strands is yet to be discovered.

1.3.2.2. Replication gradient

Due to single origin of replication, a gradient is made between early and late replicating regions in bacterial chromosomes, which affects organization and expression of genes along each replicore: usually highly expressed genes are located towards the origin whereas weakly expressed genes are located towards terminus (Rocha, 2004). This gradient across the replicores is more prominent for transcription and translation genes. The multiple gene dosage caused by multiple replication forks near the origin is primarily responsible for the high expression of genes near the origin. In *E. coli* B/r the doubling time is 20 minutes where as the replication time for the chromosomes is 45 minutes. As a result, copy number of some genes near the origin is eight times the genes near the terminus (Rocha, 2008). Schmid and Roth (1987) demonstrated the gene dosage effect by studying the expression level of *his*

operon at sixteen different locations of *Salmonella typhimurium* chromosome. Apart from the gene expression, comparison of homologous genes from *E. coli* and *S. enterica* had revealed that substitution rates in genes present near early replication regions in chromosomes (origin) is about half that of genes located towards the late replication regions in chromosomes (terminus) (Sharp *et al.*, 1989; Sharp, 1991; Mira and Ochman, 2002). The distance effect on base substitution was originally attributed to more frequent recombination repair or biased gene conversion arising from the higher gene dosage near the origin, as achieved by the presence of multiple replication forks (Sharp *et al.*, 1989; Sharp, 1991). However, further studies on this aspect had revealed that the distance effect is caused primarily by an increased rate of certain transversions near the replication terminus (Mira and Ochman, 2002; Daubin and Perriere, 2003), thereby making the terminus of a bacterial chromosome relatively enriched with 'A' and 'T' nucleotides in comparison to the origin of replication (Guindon and Perriere, 2001; Daubin and Perriere, 2003). Selection on gene orientation, length, and codon usage with respect to the position of replication origin and terminus is different, which is partly contributing to the 'A' and 'T' enrichment near the terminus (Arakawa and Tomita, 2007).

1.4. Genetic code and synonymous codon bias

Out of the possible 64 triplets from four nucleotides A, C, G, and T, 61 triplets code for 20 different amino acids in the coding region of a gene. The other three triplets (UAA, UAG, UGA) are known as stop codons signalling end of the protein synthesis. Out of the 20 amino acids, 18 (except methionine and tryptophan) are having codon degeneracy i.e. more than one codon are coding these amino acids. Codons coding the same amino acid are known as synonymous codons. The 18 amino acids are having two to six folds codon degeneracy. Though synonymous codons encode the same amino acid, they are used with different frequencies. The nonrandom usage of synonymous codons, otherwise called as codon usage bias (CUB), is common in prokaryotes, eukaryotes and viruses.

1.4.1. Earlier studies on codon usage bias

In the 1980s with increase in genome sequence database, many reports of statistical studies on codon usage in different organisms were published (Grantham *et al.*, 1980a, 1980b, 1981; Ikemura, 1981, 1982, 1985; Sharp and Li, 1986a, 1986b, 1987a, 1987b; Grosjean and

A Statistical study on the nucleotide composition of bacterial chromosomes.

Fiers, 1982; Gouy and Gautier, 1982; Bennetzen and Hall 1982; Wright 1990). Several important findings were reported in these publications which are still guiding the researches in this field. Grantham and co-workers proposed the famous 'genome hypothesis' which states "all genes in a genome, or more loosely genome type, tend to have the same coding strategy." The coding strategy in a particular genome is always conserved i.e. the use of synonymous codons has a uniformity within a genome. Different organisms are having distinct codon bias i.e. the coding strategy between organism has no similarity. Multivariate statistical technique namely correspondence analysis was used to find the codon bias in genes. Grantham and co-workers (1981) analysed thirteen highly and sixteen weakly expressed genes in *E. coli* and found difference in codon usage bias in these two types of genes. They proposed a modulation in coding strategy which states that codons found in abundant mRNA are under selection. Gouy and Gautier (1982) found that codon usage of highly expressed *E. coli* genes was different from the rest genes and the codon bias was depending on translation process (i.e. abundant tRNA). Analyzing the codon usage of three organisms, Ikemura (1981, 1982, 1985) demonstrated that in *E. coli*, *Salmonella typhimurium*, and *Saccharomyces cerevisiae* codon bias was correlated with the abundance of the cognate tRNA. Codons having abundant cognate tRNA are known as optimal codons. Bennetzen and Hall (1982) introduced the concept of codon bias index (CBI). The codon bias index is a fraction whose numerator is the total number of times that the preferred codons are used in the protein minus the random expected number of such codons. The denominator is the difference between total number of codons (excluding methionine, tryptophan and aspartic acid) and number of preferred codons expected under randomness (Bennetzen and Hall, 1982). Analysing yeast genes they found 96% percent of the 1004 amino acids were coded for by 25 preferred codons out of 61. They observed a similar phenomenon in case of highly expressed genes of *E. coli*. The expression level of a gene had a strong correlation with codon bias. These preferred codons were found having complementary bases to the most abundant tRNA isoacceptor.

1.4.2. Measures of codon usage

In silico determination of codon usage bias is a major challenge for the researchers working in this field. In the last thirty years this exercise has been done by many scientists in

different organisms with different methodologies. Measures and indices of codon usage bias based on different assumptions were developed to find out major trend of codon usage in organisms. Some of these indices are summarized here.

1.4.2.1. Fop

Ikemura (1981) introduced the concept of frequency of optimal codons. Optimal codons were originally determined for *E.coli* and *S. cerevisiae* on the basis of tRNA content and nature of codon anticodon interaction. These codons are used in highly expressed genes with maximum frequency where the codon usage bias is more. Fop is the simplest measure which is given by

$$Fop = \frac{X_{op}}{X_{op} + X_{non}}$$

Where X_{op} and X_{non} are the frequencies of optimal and nonoptimal codons in a gene 'g'. Methionine, Tryptophan and other amino acids whose optimal codons are not known are excluded from the calculation.

1.4.2.2. P2

P2 index of codon usage was developed by Gouy and Gautier (1982). It is the proportion of codons which conform to codon anticodon interaction rule. P2 is given by

$P2 = (WWC+SSU)/(WWY + SSY)$ where W= A or U, S= G or C, Y= C or T.

1.4.2.3. RSCU

The relative synonymous codon usage (RSCU) (Sharp *et al.* 1986) for each codon is calculated as the observed number of occurrences divided by the number expected if all the synonymous codons for an amino acid were used equally. For synonymous codon i of an k-fold degenerate amino acid is given by

$$RSCU = \frac{x_i}{\frac{1}{k} \sum_{i=2}^k x_i}$$

Where X_i is the number of occurrences of codon i, k is 1,2,3,4 or 6.

1.4.2.4. CAI

The codon adaptation index (CAI) (Sharp and Li, 1987b) measures the unidirectional codon usage bias in a gene. CAI estimates the extent to which codons of a gene are adapted towards the optimal codons favored by highly expressed genes. The relative adaptedness (w_i) of a codon i is measured from the RSCU values of the codons obtained from a set of highly expressed genes.

$$w_i = \frac{RSCU_i}{RSCU_{max}} = \frac{x_i}{x_{max}}$$

Where RSCU and X values are considered from a reference set of highly expressed genes. The CAI for a gene 'g' is defined as the geometric mean of w values for codons in that particular gene and is given by

$$CAI = \left(\prod_{i=1}^L w_i \right)^{\frac{1}{L}}$$

L is the number of codons in that particular gene excluding methionine, tryptophan and stop codons.

1.4.2.5. ENc

Wright (1990) introduced the measure of effective number of codons (ENc) for a gene. It is a general measure of bias in a gene from equal usage of alternative synonymous codons. It reaches its maximum value 61 when all the synonymous codons are equally used in a particular gene. Its minimum value is 20 which is obtained when a gene uses only one codon per amino acid. Knowledge of optimal codon or the reference set of gene are not required in calculating ENc. For a particular gene ENc is given by

$$ENc = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}$$

Where F_k is the average frequency of k fold degenerate amino acids.

F_k for each of k fold degenerate amino acid is given by

$$F_k = \frac{n \sum_{i=1}^k \left(\frac{n_i}{n} \right)^2 - 1}{n-1}$$

Where n is the total number of codons for that amino acid, n_i is the number of occurrence of i^{th} codon for this amino acid.

ENc value is affected by silent GC content θ_g of a gene 'g', an equation to approximate the relationship under the hypothesis of no selection was proposed by Wright (1990) which is given by

$$f(\theta_g) = 2 + \theta_g + \frac{29}{(\theta_g)^2 + (1 - \theta_g)^2}$$

Wright suggested the use of ENc plot where ENc values are plotted against θ_g with the curve $f(\theta_g)$ is superimposed on it. This was a part of the strategy to investigate mutational bias in synonymous codon usage. If the ENc values progress along the side of the curve, then it is an indication of significant mutational bias in the codon usage.

1.4.2.6. Shannon information based codon bias

Zeeberg (2002) developed a method based on Shannon information theory to compute synonymous codon usage bias in coding regions of different organism. The information measure for a given sequence 's', which is a member of a set 'G' is given by the uncertainty difference in 's' and 'G'. The general uncertainty measure was defined by

$$Uncertainty = H = \sum_{i=1}^{n_{aa}} \left(\sum_{j=1}^{n_{\text{syncod}(i)}} p_{i,j} \log_2(p_{i,j}) \right)$$

Where n_{aa} is the effective number of amino acids (here n_{aa} is taken as 23 counting leu₂, ser₂ and arg₂ as separate amino acid), $n_{\text{syncod}(i)}$ is the number of synonymous codons for the i^{th} amino acid and $p_{i,j}$ is the probability that amino acid i will be coded by its j^{th} synonymous codon. The information for the reference sequence s is

$$Information = H_g - H_s$$

In the next phase using a theoretical model the information function is to be estimated (Zeeberg, 2002).

1.4.2.7. ICDI

The Intrinsic codon deviation index (ICDI) developed by Freire-Picos *et al.*, (1994) is based on RSCU values of the 18 amino acids with at least two fold codon degeneracy. The ICDI value for a gene is expressed in terms of s_k values given by

$$s_k = \sum \frac{(n_i - 1)^2}{k(k-1)}$$

Where n_i is the RSCU value for the i^{th} codon and k is the codon degeneracy number, ($k=2,3,4$ and 6). The ICDI is given by

$$ICDI = \sum s_2 + s_3 + \sum s_4 + \frac{\sum s_6}{18}$$

A gene with strong codon bias will have maximum ICDI value.

1.4.2.8. tAI_g

dos Reis *et al.* (2004) introduced tRNA adaptation index (tAI_g) keeping aim at speculating translational selection in a gene 'g' with the help of its tRNA usage. CAI is a measure of unidirectional codon usage bias giving relative measure of codon adaptation towards optimal codons used by a reference set of genes. Similarly tAI_g is giving a measure of how well the gene in question adapted towards tRNA gene pool of a genome. tAI_g is calculated with the help of absolute adaptiveness value W_i for each codon i which is given by

$$W_i = \sum_{j=1}^{n_j} (1 - s_{ij}) GCN_j$$

n_j is the number of tRNA isoacceptor for the i th codon, $tGCN_{ij}$ is the gene copy number of the j th tRNA that recognizes the i th codon. s_{ij} is the selective constraint on the efficiency of codon anticodon coupling. The relative adaptiveness value ω_i is given by

$$w_i = \frac{W_i}{W_{\max}} \text{ if } W_i \neq 0,$$

$$\text{or } = W_{\text{mean}} \text{ else}$$

where W_{\max} and W_{mean} are the maximum and average values of W_i . tAI_g of a gene 'g' is defined by

$$tAI_g = \left(\prod_{k=1}^{l_g} w_{i_{k_g}} \right)^{\frac{1}{l_g}}$$

Where i_{k_g} is the codon defined by k th triplet in a gene g . l_g is the codon length of the gene 'g'.

1.4.2.9. Correspondence Analysis

Correspondence analysis (Benzecri, 1973) is one of the oldest multivariate statistical methods used in codon usage analysis. Grantham *et al.* (1980a, 1980b) has used correspondence analysis to study codon usage in the genes of different organisms. This method has been found most widely used by many scientists in the codon usage studies. This method has flexibility in accommodating large set of codon usage data presented in contingency table. In general, data for correspondence analysis are presented in the form of relative codon usage rather than the absolute codon counts. Codon count data are generally not used to avoid bias due to amino acid usage. It is a powerful tool to find major trends in the data. It isolates the major trends amidst stochastic noise. The disadvantage of this method is that it provides no interpretation of the available trends in the data. In correspondence analysis of codon usage in the genes of an organism each gene is first plotted as a point in the multidimensional space with 61 coordinates. In the next phase the points are projected to a lower dimensional space whose first two axes correspond to most important variations in codon usage (Grantham *et al.*, 1980). Genes having similar strategy for codon usage are grouped together by the correspondence analysis. The grouping is carried out with the help of

perpendicular distance between them in the multidimensional space of 61 axes. The first axis generated by the correspondence analysis is known as the Principal axis and it gives the coordinates of the genes with respect to the major source of variation.

1.4.3. Underlying hypotheses for codon usage bias

Codon usage bias in organisms has been studied with respect to two hypotheses – hypothesis based on selection and hypothesis based on mutation (or neutralist point of view) (Hershberg and Petrov, 2008).

1.4.3.1. Hypothesis based on natural selection

The hypothesis based on natural selection explains the codon usage bias in the light of efficiency and accuracy of protein synthesis. Genes expressed at high level use a preferred set of codons having complementary bases to the most abundant tRNA isoacceptor. Consequently biased codon usage in highly expressed genes has been explained by translational selection (Ikemura, 1981; Bennetzen and Hall, 1982; Guoy and Gautier, 1982). The high correlation between the abundance of *E.coli* tRNA and the frequency of respective codons (Ikemura, 1981) proves the role of translational selection in codon usage bias. Experiment in case of yeast genes (Bennetzen and Hall, 1982) revealed the similar result. Codon usage bias of similar kind has been found in *Drosophila melanogaster* (Shields *et al.* 1988). Nucleotide substitution rates are less in case of highly expressed genes than that of weakly expressed genes (Sharp and Li, 1987a).

1.4.3.2. Hypothesis based on mutation

The hypothesis based on mutation (more generally the neutralist point of view) believes that codon usage bias in the coding regions of the genomes is the outcome of nonrandom mutations. From the neutralist point of view, the significant parameter explaining codon usage bias is the GC content of a genome which is believed to be maintained by mutational processes. Mutation generated codon usage bias has been studied from two aspects- (i). genomic G+C content (Muto and Osawa, 1987; Chen *et al.*, 2004), (ii). strand-specific mutational bias (Lobry, 1996; McInerney, 1998; Frank and Lobry, 1999). Using correspondence analysis, McInerney (1998) had shown that strand specific mutational bias was the major source of codon usage bias in *B. burgdorferi*. The major trend of codon usage

A Statistical study on the nucleotide composition of bacterial chromosomes.

in *B. burgdorferi* have separated the genes in two parts – transcribed in the leading strand and transcribed in the lagging strand (McInerney, 1998).

1.4.4. The Selection-Mutation-Drift (SMD) theory and population genetics model

The selection-mutation-drift theory explains the pattern of synonymous codon usage in a finite population as the resultant effect of three forces- the natural selection favouring optimal codons for efficient and accurate translation of the protein product, non-random mutations and random genetic drift allowing non-optimal codons (Li, 1987; Shields, 1990; Bulmer, 1991). The findings of the aforementioned studies give sufficient evidences to believe that codon usage spectra of different organisms are marked by both selection and mutation. The intensity of selection and mutation may vary from organism to organism. Population genetics model was used to measure the intensity of selection influencing codon usage bias. The model developed by Bulmer (1991) estimates selection taking into account the codon usage bias of those amino acids having only one optimal codon. Considering a haploid population using two alleles B_1 and B_2 with relative frequencies p_t and q_t at time t with fitness coefficients 1 and $1-s$ ($s > 0$) the equilibrium gene frequency P was found to satisfy the following equation-

$$sP(1-P) + v(1-P) - uP = 0 \quad (\text{Bulmer, 1991})$$

where u is the mutation rate from B_1 to B_2 and v is the rate from opposite direction. Instead of considering P as a constant value it was considered as a random variable having probability density function

$$f(p) \propto e^{Sp} p^{V-1} (1-p)^{U-1}$$

where

$$S = 2Nes, V = 2Nev, U = 2Nev, Ne = \text{effective population size}$$

In a small population where $U+V \ll 1$ it is expected to see monomorphism with a fraction P for B_1 and $(1-P)$ for B_2 where P is given by

$$P = e^S V / (e^S V + U) \quad (\text{Bulmer, 1991})$$

Sharp *et al.* (2005) has utilized this model in estimating S from the above equation

$$S = \ln(P.k / (1 - p)) \quad \text{where } k = U/V.$$

When selection for codon bias is too small i.e. $S \rightarrow 0$ then $e^S \rightarrow 1$, as a result in genes with weak selection

$$P = V / (V + U)$$

Now it follows $k = (1 - P) / P$, using $k = U/V$

Now S can be estimated putting $k = (1 - P) / P$.

Using this principle Sharp *et al.* (2005) has estimated the strength of selected codon usage bias in 80 bacterial genomes. Codon usage bias in four amino acids namely phenylalanine, tyrosine, isoleucine and asparagine in weakly expressed genes was used in estimating k since in all the species C ending codons for these four amino acids were found optimal. Isoleucine was considered as a two codon amino acid neglecting rare codon AUA.

Higgs and Ran (2008) has used a modified version of this model using GC content θ as an additional parameter and selected codon usage bias S was presented in terms of codon counts in highly and weakly expressed genes which is given by

$$S = \ln \left(\frac{n_C^{high} n_U^{low}}{n_C^{low} n_U^{high}} \right)$$

Similar relation for A and G ending codon families was also derived with subscript G replacing C and A replacing U. Using these models on five organisms from prokaryotes to eukaryotes the authors have shown that strength of selection obtained using C, U ending codons in some cases opposite to that of result obtained by using A, G ending codons. Their argument in this respect is that the coevolution of codon usage and tRNA gene content may show different stable state of codon usage in the same organism. Considering translational

kinetics the authors have developed another model showing relation between strength of selected codon usage bias S and the relative rates of translation b_{xy} . S was shown depending on b_{xy} s through a constant K , the cost of translation in an organism. The authors have elegantly shown that species with significant translational selection may have alternative stable states of codon usage.

1.5. Discussion

Nucleotide composition in genomes is found to be marked by the phenomena described above. These studies have been of major interest to evolutionary biologists in the light of two theories of evolution i.e. selection vs. mutation.

Chargaff's 2nd parity is observed in many chromosomes. This indicates that the feature is under selection in these genomes. It is important to study intra-strand parity in chromosomes with respect to different oligonucleotides and compare the magnitude of parity violations. The methods adopted by different authors described in this chapter to study PR2 do not give freedom to study parity violation with respect to an oligonucleotide and its complement, which will be an important way to answer the evolutionary significance of PR2 in genomes. In Chapter II we have described a method which is useful in this respect.

The neutral theory of evolution was proposed by Kimura in 1968 and by Jukes and Kings in 1969. The degeneracy of codons in the genetic code and the neutral theory of evolution complemented well. However, the seminal findings of Ikemura in *E. coli* and *S. cerevisiae* provided vital support to selectionists' view of codon usage in genomes. Evolution of genome GC% in bacteria is one of the common examples of neutral theory of evolution. The recent discovery of SSMB in bacterial chromosomes and its influence on codon usage is an interesting case to test the above two theories of evolution. According to the selection-mutation-drift (SMD) theory, selection is a dominant factor over mutation in an organism. It will be interesting to study the SSMB influence on codon usage with respect to SMD. This has been discussed in Chapter III. The observation of Sharp *et al* (2005) suggests that force of selection varies among different genomes. However, it is not mentioned directly in their study whether the variation in selection is dependent or is independent of mutational bias. A genome might exhibit high selection due to low mutation or the *vice versa*. This question has been addressed in Chapter IV.

CHAPTER II

2. A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (Chargaff's 2nd parity rule)

2.1. Abstract

Chargaff's rule of intra-strand parity (ISP) between complementary mono/oligo nucleotides in chromosomes is well established in the scientific literature. Although a large numbers of papers have been published citing works and discussions on ISP in the genomic era, scientists are yet to find all the factors responsible for such a universal phenomenon in the chromosomes. In the present work, this issue has been addressed from a new perspective, which is a parallel feature to ISP. The compositional abundance values of mono/oligo-nucleotides were determined in all non-overlapping sub-chromosomal regions of specific size. Also the frequency distributions of the mono/oligo-nucleotides among the regions were compared using Kolmogorov-Smirnov test (KS-test). Interestingly, the frequency distributions between the complementary mono/oligo-nucleotides revealed statistical similarity, which we named as intra-strand frequency distribution parity (ISFDP). ISFDP was observed as a general feature in chromosomes of bacteria. Violation of ISFDP was also observed in several chromosomes. Chromosomes of different strains belonging to a species in bacteria (*H. influenza*, *X. fastidiosa*, etc) are found to be different among each other with respect to ISFDP violation. ISFDP correlates weakly with ISP in chromosomes suggesting that the latter one is not entirely responsible for the former. Asymmetry of replication topography and composition of forward encoded sequences between the strands in chromosomes are found to be insufficient to explain ISFDP feature in all chromosomes. This suggests that multiple factors in chromosomes are responsible for establishing ISFDP.

2.2. Introduction

Chargaff's 1st parity rule based on nucleotide composition of double stranded DNA states that the complementary nucleotides have the same abundance values (Chargaff, 1950, 1951; Forsdyke and Mortimer, 2000). This is explained by the DNA double helix model in which A pairs only with T, and G pairs only with C (Watson and Crick, 1953). Chargaff and his colleague came with a similar observation of compositional relationship between complementary nucleotides even within individual DNA strands of bacterial chromosomes

(Rudner *et al.*, 1968; Rudner *et al.*, 1969). In the post genomic era, this intra-strand relationship between complementary nucleotides is observed in double stranded genomes of viruses, bacteria, archaea and eukaryotes, which is known as Chargaff's 2nd parity rule or intra-strand parity (ISP) (Forsdyke and Mortimer, 2000). There is no such defined rule to describe ISP in chromosomes like the base pairing rule in Chargaff's first parity. ISP is also observed between complementary oligonucleotides in chromosomes (Prabhu, 1993; Qi and Cuticchia, 2001; Baisnée *et al.*, 2002; Verma *et al.* 2005), which has been attributed to genome wide large scale inversion, inversion transposition (Albrecht Buehler, 2006) and coding sequence compositional symmetry between the strands (Verma *et al.*, 2005). Violation of ISP is observed with respect to organellar (mitochondria and plastids) genomes of some organisms, single stranded viral genomes or any RNA genome (Mitchell and Bridge, 2006; Nikolaou and Almirantis, 2006; Deng, 2007).

Theoretically, under no strand bias in terms of mutation and selection, the base complementary relationship easily explains the presence of ISP in chromosomes (Sueoka, 1995; Lobry, 1995). However, several evidences now prove that both the strands are not identical in terms of mutation/selection (Frank and Lobry, 1999). This results into violation of ISP in sub-chromosomal regions. Longer the sub-chromosomal region smaller is the violation of ISP observed (Nikolaou and Almirantis, 2005). The mechanisms that are responsible to cause the violation are defined under three categories (Sueoka, 1999). Firstly, DNA replication: leading strand (LeS) is found to be composed of more K nucleotides (G & T) than the complementary M (A & C) nucleotides and the reverse holds true for the lagging strand (LaS) (Rocha *et al.*, 1999). This is due to the fact that the LeS which functions as the template for Okazaki fragment synthesis (functions as template for LaS) remains exposed more as single stranded than the LaS (functions as template for LeS) during replication that results into higher deamination of the cytosine residues (Lobry and Sueoka, 2002; Grigoriev, 1998) in LeS (cytosine gets deaminated 140 times faster in ssDNA than in dsDNA; Francino & Ochman, 1997). In addition, the influence of Okazaki fragments and the sliding DNA clamp proteins associated with the synthesis of LaS create functional asymmetry of the mismatch repairing system on DNA (Francino and Ochman, 1997). Secondly, transcription: genes are preferentially located in the LeS than the LaS to avoid head on collision between

the machineries of replication and transcription (Johnson and O'Donnell, 2005). During transcription, the non-template strand remains more exposed as single stranded than the template strand, which causes asymmetry in cytosine deamination between the strands (Bell and Forsdyke, 1999). The transcription coupled repair system also acts only upon the template strand and thereby contributes to the strand asymmetry (Francino *et al.*, 1996). Thirdly, translation: uses of synonymous codons are influenced by differential abundance of tRNA molecules which results into the differential abundance of complementary nucleotides at the 3rd position of family box codons. This causes parity violation (Sueoka, 1995). In spite of these factors favouring violations of the parity in chromosomes, ISP is observed in an entire chromosome due to the cancellation effect of the local violations in opposite directions (Sueoka, 1995).

Evolutionary biologists are more interested to understand the role of mutation and/or selection in the violation of ISP by analyzing the weakly selected or selectively neutral regions (3rd position of family box codons and non coding regions) in chromosomes (McLean *et al.*, 1998; Sueoka, 1995). Whether any specific feature(s) is/are associated with chromosomes exhibiting ISP is yet to be understood. Shioiri and Takahata (2001) studied ISP by finding out total AT skew (ATS) and GC skew (GCS) in the chromosomes of several bacteria. In their study, out of 36 bacterial chromosomes, *Xylella fastidiosa* exhibited maximum ATS and GCS (Shioiri and Takahata, 2001). They observed variable ATS/GCS among chromosomes of different strains of a species as well as chromosomes within a bacterial cell. They also observed ATS and GCS may be different from each other within a chromosome. Since, they did not do any statistical analysis of the skew, the significance of the variability observed among chromosomes were not discussed by them. The usual statistical tool used to find out ISP in chromosomes is a correlation analysis of oligonucleotides abundance described by Prabhu (1993). ISP study between complimentary mononucleotides is important because it has been proven that oligo-nucleotide parity and mononucleotide parity are independent (Baisnée *et al.*, 2002). Baisnée *et al.*, (2002) studied parity in chromosomes by measuring the S^1 index which is defined as the sum of the absolute values of the differences between complementary oligonucleotides (n mer) frequencies (n varies from 1 – 9 mer). Both these methods do not measure the statistical significance of

differences between the abundance values of a mono/oligonucleotide and its reverse complement. For example, if a chromosome carries significant similarity between the abundance values of A and T but carries significant difference between the abundance values of G and C, this will not be identified separately. Similarly, the above methods are unable to find out parity violations in chromosomes with respect to the abundance values of an oligonucleotide and its reverse complement. We have developed a methodology here that can independently study ISP between S nucleotides (any oligonucleotide and its reverse complement) as well as between W nucleotides using the abundance values of mononucleotides. We use the well known Kolmogorov-Smirnov test to study the frequency distribution of the compositional abundance values of the mononucleotides in a chromosome sequence, which gives the statistical significance of the similarity between the distributions of complementary nucleotides. This we called as intra-strand frequency distribution parity (ISFDP), which has been used here to study chromosomes of bacteria.

2.3. Materials and Methods

2.3.1. Frequency distribution calculation

Chromosome sequences of different bacteria (Table 2.1) were obtained from genome information broker, DDBJ site (www.gib.genes.nig.ac.jp). Bacterial chromosomes were chosen randomly from the database starting the genus name from A to Z. Chromosome sequences of different strains belonging to same species in case of bacteria were taken in several cases to do intra-species comparison. Each chromosome sequence was divided into smaller size sequences of 1000 nucleotides each starting from the beginning and the abundance value of the four nucleotides were determined using a computer program (developed for this study). The distribution of the abundance values of complementary nucleotides in different fragments were analyzed by Kolmogorov-Smirnov non parametric test using XLSTAT package (Kovach computing services, Anglesey, Wales). H_0 : distribution patterns of any two nucleotides/oligonucleotides in a chromosome are similar; H_A : there is difference between the two distributions. Due to large sample size, similarity was considered at the p value > 0.01 , weak similarity was considered at the p value between 0.01 and 10^{-4} and the value $< 10^{-4}$ was considered as strong violation similarity. Group frequency distribution of the abundance values were plotted to observe the frequency distribution parity.

In case of the di and tri nucleotides, the abundance values were determined using a different computer program (developed here for this study) in the segments for the 16-dinucleotides and 64 trinucleotides, respectively. The analysis was done as described for the mononucleotides above.

Angular replication asymmetry of the chromosomes was calculated with the help of the information on *ori* (origin) and *ter* (termination) cited in the web sites (<http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/origin/>; <http://pbil.univ-lyon1.fr/software/Oriloc/oriloc.html>; Frank and Lobry, 2000). The chromosomal region starting from *ori* to the *ter* was considered as the leading region in the Watson strand (Ws) and the remaining portion of the chromosome as the lagging region. For a circular chromosome the angular replication asymmetry was calculated as the amount of angular distance of leading region deviating from 180°.

2.3.2. Proportionate distribution of forward encoded and reverse encoded sequences in a DNA strand

From the DDBJ site only coding sequences were downloaded. A continuous stretch of the nucleotide sequence was made from all the sequences by removing the gene names. This resembled a DNA strand only composed of forward encoded sequences. Frequency distribution analysis was done on this. In another approach, 50% of the above strand was made reverse complement by *in silico* followed by joining with the rest. This resembled a DNA strand composed of 50% forward encoded and 50% reverse encoded sequences. Frequency distribution study was carried out as described above.

2.3.3. Calculation of whole genome AT skew and GC skew

$ATS = |(\sum A - \sum T)| / (\sum A + \sum T)$ and $GCS = |(\sum G - \sum C)| / (\sum G + \sum C)$ Where $\sum A$ is the total of A in all the windows of a chromosome and similar definition for other three sums.

2.3.4. Identification of leading and lagging strand region

AT and GC skew analysis of the chromosome sequences were done as described earlier. Cumulated AT skew and GC skew plots were used to find out the tentative leading and lagging portions in a DNA strand.

2.3.5. Relative proportion of coding sequence distribution

This was found out by deducting ORF numbers between Watson strand (Ws: top strand) and Crick strand (Cs: bottom strand) followed by dividing that with total number of ORFs. Gene orientation information was obtained from the web site (<http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi>).

2.4. Results

2.4.1. Intra-strand frequency distribution parity in chromosomes of bacteria

In this study a total of 112 bacterial chromosomes were considered, which includes different lineages of bacteria such as protobacteria, cyanobacteria, firmicutes, actinobacteria etc. Samples from each group were taken randomly. The bacteria included in the sample comprised of a GC% variation from a minimum of 28% to a maximum of 75%, and chromosome size variation from 580 kb to a maximum of 9105 kb. We have studied the frequency distributions of the abundance values of mono nucleotides in the uniform sub-chromosomal length of 1000 nucleotides. A collective analysis of the nucleotide abundance values from all the segments of a chromosome was done by frequency distribution smooth curves using Microsoft Excel and the similarity of the distributions of two complementary nucleotides was tested using Kolmogorov-Smirnov test (XL-Stat; <http://www.xlstat.com/en/download>). Fig. 2.1 {a(i), b(i), c(i), d(i), e(i)} represent the smooth curves of frequency distributions of nucleotides in chromosomes *Campylobacter jejuni* RM1221 (30.31%), *Escherichia coli* K12 MG1655 (50.79%), *Xanthomonas campestris* pv. *campestris* (Xcc;65.07%), *Xylella fastidiosa* 9a5c (52.68%), and *Xylella fastidiosa* Temecula (51.78%). Smooth curves of complementary nucleotides overlap with each other in the first three chromosomes while that of non-complementary ones do not. In the fourth chromosome none of the curves overlap with each other. In *E. coli* chromosome {Fig. 2.1 b(i)} all the four smooth frequency curves are close to each other due to the closeness of the abundance values of the nucleotides whereas in the graphs of *C. jejuni* and *Xcc* the smooth frequency curves of W (A & T) and S (G & C) nucleotides are distinctly separated as GC% the chromosome are towards both extremes. The distribution was studied by Kolmogorov-Smirnov test (KS test) and the results of the five chromosomes are shown in Fig. 2.1 {a (ii, iii), b (ii, iii), c (ii, iii), d

(ii, iii), & e (ii, iii)}. The graphs generated by KS-test suggest the complete overlapping between the complementary nucleotides in the chromosomes except the one of *X. fastidiosa* strain, which is in concordant with the smooth frequency curves. The distributional similarity between complementary nucleotides is called as intra-strand frequency distribution parity (ISFDP). A total of 112 bacterial chromosomes (Table 2.1) were analyzed by the KS-test to study ISFDP. The p-values between the A and T distributions as well as between the G and C distributions are given (Table 2.1).

Table 2.1: Result of Kolmogorav-Smirnov (KS) test for significance between the frequency distribution of complementary nucleotides.

Sl no	Strain name	Size (Kb)	GC%	KS (W)	KS(S)	$ \sum A - \sum T / (A+T)$	$ \sum G - \sum C / (G+C)$	Bacterial Group	TB (in°)
1	<i>Acinetobacter sp.</i> ADP1	3598	40.43	0.745	0.006	0.00068	0.00484	γ -Proteobacteria	7.07
2	<i>Actinobacillus pleuropneumoniae</i> L20 serotype 5b	2274	41.3	0.436	0.819	0.00187	0.00109		NA
3	<i>Actinobacillus succinogenes</i> 130Z	2319	44.91	0.312	0.291	0.00232	0.00291		
4	<i>Aeromonas hydrophila subsp. hydrophila</i> ATCC 7966	4744	61.55	0.88	0.19	0.00141	0.00139		
5	<i>Aeromonas salmonicida subsp. salmonicida</i> A449	4702	58.51	0.04	0.959	0.00215	0.00073		
6	<i>Agrobacterium tumefaciens</i> C58 (circular chromosome)	2841	59.38	< 0.0001	< 0.0001	0.00694	0.00967	α -Proteobacteria	7.37
7	<i>Alkaliphilus oremlandii</i> OhILAs	3123	36.26	< 0.0001	< 0.0001	0.00615	0.01324	Firmicutes	NA
8	<i>Anaeromyxobacter dehalogenans</i> 2CP-	5013	74.9	0.077	0.001	0.00476	0.00249	δ -Proteobacteria	70.57

C								
9	<i>Anaeromyxobacter</i> sp. Fw109-5	5277	73.53	0.712	0.008	0.00073	0.00216	7.48
10	<i>Bacillus anthracis</i> Ames	5227	35.38	0.004	< 0.0001	0.00215	0.00581	NA
11	<i>Bacillus anthracis</i> 'Ames Ancestor'	5227	35.38	0.003	< 0.0001	0.00215	0.00582	7.48
12	<i>Bacillus anthracis</i> Sterne	5228	35.38	0.008	< 0.0001	0.00221	0.00588	7.46
13	<i>Bacillus subtilis</i>	4214	43.52	0.219	0.234	0.00212	0.00224	13.69
14	<i>Bacillus</i> <i>thuringiensis</i> Al Hakam	5257	35.43	0.123	0.002	0.00042	0.00081	NA
15	<i>Bacillus</i> <i>thuringiensis</i> serovar konkukian 97-27	5237	35.41	0.015	< 0.0001	0.00194	0.00438	3.98
16	<i>Bordetella</i> <i>parapertussis</i> 12822	4773	68.1	0.433	< 0.0001	0.00247	0.00776	37.01
17	<i>Bordetella pertussis</i> Tohama I	4086	67.72	0.861	< 0.0001	0.00022	0.00390	71.28
18	<i>Bradyrhizobium</i> <i>japonicum</i> USDA 110	9105	64.06	0.512	0.31	0.00070	0.00038	7.07
19	<i>Bradyrhizobium</i> sp. BTAi1	8264	64.92	0.381	0.01	0.00100	0.00163	NA
20	<i>Brucella melitensis</i> 16M	1177	57.35	0.472	0.008	0.00227	0.00312	
21	<i>Campylobacter</i> <i>concisus</i> 13826	2052	39.43	0.033	0.048	0.00038	0.00599	
22	<i>Campylobacter</i> <i>curvus</i> 525.92	1971	44.54	0.028	0.752	0.00745	0.00282	
23	<i>Campylobacter</i> <i>jejuni</i> RM1221	1777	30.31	0.574	0.23	0.00330	0.00436	8.69
24	<i>Campylobacter</i> <i>jejuni</i> subsp. <i>jejuni</i>	1628	30.54	0.491	0.029	0.00250	0.00613	NA

A Statistical study on the nucleotide composition of bacterial chromosomes.

	81116								
25	<i>Campylobacter jejuni</i> subsp. <i>jejuni</i> NCTC 11168	1641	30.55	0.067	0.132	0.00296	0.00457		10.25
26	<i>Candidatus Desulfococcus oleovorans</i> Hxd3	3944	56.17	0.258	0.133	0.00199	0.00157	Firmicutes	NA
27	<i>Caulobacter crescentus</i> CB15	4016	67.22	0.042	0.171	0.00396	0.00188	α -Proteobacteria	8.56
28	<i>Chlamydia muridarum</i> Nigg	1072	40.34	0.221	0.853	0.00107	0.00337		1.17
29	<i>Chlamydia trachomatis</i> AHAR-13	1044	41.31	0.228	0.284	0.00230	0.00059	Chlamydiae	1.30
30	<i>Chlamydophila abortus</i> S263	1144	39.87	0.534	0.002	0.00065	0.00361		0.57
31	<i>Coxiella burnetii</i> Dugway 7E9-12	2158	42.44	0.004	0.001	0.00592	0.00573	γ -Proteobacteria	NA
32	<i>Coxiella burnetii</i> RSA 493	1995	42.66	0.014	0.467	0.00198	0.00029		31.15
33	<i>Desulfovibrio desulfuricans</i> G20	3730	57.84	0.59	0.001	0.00189	0.00322	Firmicutes	10.70
34	<i>Desulfovibrio vulgaris</i> subsp. <i>vulgaris</i> DP4	3462	63.01	0.3	0.159	0.00152	0.00106		NA
35	<i>Desulfovibrio vulgaris</i> subsp. <i>vulgaris</i> Hildenborough	3570	63.14	0.557	0.082	0.00143	0.00024	δ -Proteobacteria	4.78
36	<i>Enterobacter sakazakii</i> ATCC BAA-894	4368	56.77	0.167	0.388	0.00359	0.00044		NA
37	<i>Enterobacter</i> sp. 638	4518	52.98	0.645	0.39	0.00169	0.00163	γ -Proteobacteria	NA
38	<i>Escherichia coli</i> 536	4938	50.52	0.714	0.084	0.00062	0.00328		7.40
39	<i>Escherichia coli</i>	5082	50.55	0.779	0.576	0.00032	0.00070		NA

A Statistical study on the nucleotide composition of bacterial chromosomes.

APEC O1								
40	<i>Escherichia coli</i> CFT073	5231	50.48	0.112	0.92	0.00173	0.00080	5.66
41	<i>Escherichia coli</i> E24377A	4979	50.62	0.736	0.128	0.00205	0.00212	NA
42	<i>Escherichia coli</i> HS	4643	50.82	0.328	0.469	0.00151	0.00207	
43	<i>Escherichia coli</i> K12 MG1655	4639	50.79	0.732	0.587	0.00054	0.00113	4.28
44	<i>Escherichia coli</i> UTI89	5065	50.6	0.51	0.237	0.00076	0.00203	3.70
45	<i>Escherichia coli</i> W3110	4646	50.8	0.873	0.729	0.00073	0.00091	12.64
46	<i>Frankia alni</i> ACN14A chromosome	7497	72.82	0.463	0.036	0.00141	0.00139	Actinobacteria NA
47	<i>Frankia sp.</i> CcI3	5433	70.08	0.808	0.662	0.00129	0.00017	
48	<i>Haemophilus</i> <i>influenzae</i> 86- 028NP	1914	38.16	0.886	0.654	0.00089	0.00044	γ-Proteobacteria
49	<i>Haemophilus</i> <i>influenzae</i> PittEE	1813	38.04	0.544	0.038	0.00054	0.00317	
50	<i>Haemophilus</i> <i>influenzae</i> PittGG	1887	38.01	0.125	< 0.0001	0.00005	0.01016	
51	<i>Haemophilus</i> <i>influenzae</i> Rd KW20	1830	38.15	0.154	0.004	0.00298	0.00472	46.61
52	<i>Helicobacter</i> <i>acinonychus</i> Sheeba	1553	38.18	0	0.596	0.00869	0.00164	ε-Proteobacteria NA
53	<i>Helicobacter</i> <i>hepaticus</i> ATCC 51449	1799	35.93	0.161	< 0.0001	0.00499	0.01518	46.54
54	<i>Helicobacter pylori</i> J99	1643	39.19	0.246	0.256	0.00259	0.00510	10.97
55	<i>Lactobacillus</i> <i>acidophilus</i> NCFM	1993	34.72	0.382	< 0.0001	0.00066	0.01644	Firmicutes 19.54
56	<i>Lactobacillus brevis</i>	2291	46.22	0.023	< 0.0001	0.00271	0.02882	NA

A Statistical study on the nucleotide composition of bacterial chromosomes.

	ATCC 367								
57	<i>Lactobacillus delbrueckii</i> subsp. <i>bulgaricus</i> ATCC BAA-365	1856	49.69	0.491	0.264	0.00201	0.00087		
58	<i>Lactobacillus reuteri</i> F275	1999	38.87	0.001	< 0.0001	0.00122	0.01040		
59	<i>Lactococcus lactis</i> subsp. <i>cremoris</i> MG1363	2529	35.75	0.233	0.056	0.00352	0.00524		NA
60	<i>Lactococcus lactis</i> subsp. <i>cremoris</i> SK11	2438	35.86	0.399	0.521	0.00147	0.00136		
61	<i>Magnetococcus</i> sp. MC-1	4719	54.17	0.001	< 0.0001	0.00490	0.01198	Magnetococcus	
62	<i>Magnetospirillum magneticum</i> AMB-1	4967	65.09	0.031	< 0.0001	0.00339	0.00288	α -Proteobacteria	2.14
63	<i>Methylobacillus flagellatus</i> KT	2971	55.72	0.03	0.916	0.00226	0.00135	β -Proteobacteria	10.57
64	<i>Methylococcus capsulatus</i> Bath	3304	63.59	0.145	0.004	0.00150	0.00287	γ -Proteobacteria	NA
65	<i>Mycobacterium leprae</i> TN	3268	57.8	0.003	< 0.0001	0.00378	0.00609		7.04
66	<i>Mycobacterium</i> sp KMS	5737	68.44	0.389	0.478	0.00030	0.00060	Actinobacteria	NA
67	<i>Mycobacterium tuberculosis</i> F11	4424	65.62	0.366	0.007	0.00006	0.00198		
68	<i>Mycobacterium ulcerans</i> Agy99	5631	65.47	< 0.0001	< 0.0001	0.00433	0.00374		
69	<i>Mycoplasma gallisepticum</i> R	996	31.45	0.18	0.615	0.00626	0.00021		9.32
70	<i>Mycoplasma genitalium</i> G37	580	31.69	0	0.148	0.01219	0.00433	Tenericutes	3.75
71	<i>Mycoplasma hyopneumoniae</i> J	897	28.52	0.033	0.599	0.01020	0.00067		NA
72	<i>Mycoplasma</i>	816	40.01	0.001	0.115	0.01767	0.00243		16.23

A Statistical study on the nucleotide composition of bacterial chromosomes.

	<i>pneumoniae</i> M129								
73	<i>Neisseria gonorrhoeae</i> FA 1090	2153	52.69	0.07	0.033	0.00601	0.00144	β -Proteobacteria	9.20
74	<i>Neisseria meningitidis</i> MC58	2273	51.52	0.695	0.004	0.00135	0.00806		NA
75	<i>Nitrobacter hamburgensis</i> X14	4406	61.72	0.332	0.53	0.00112	0.00041	α -Proteobacteria	
76	<i>Nitrobacter winogradskyi</i> Nb-255	3402	62.05	0.011	< 0.0001	0.00323	0.00294		37.15
77	<i>Nitrosococcus oceani</i> ATCC 19707	3481	50.32	0.02	0.056	0.00530	0.00243	γ -Proteobacteria	8.39
78	<i>Nitrosomonas eutropha</i> C91	2661	48.49	0.992	0.318	0.00043	0.00162	β -Proteobacteria	NA
79	<i>Nostoc sp.</i> PCC 7120	6413	41.35	0.134	0.857	0.00129	0.00162	Cyanobacteria	
80	<i>Pseudomonas entomophila</i> L48 chromosome	5888	64.16	0.657	0.251	0.00078	0.00173	γ -Proteobacteria	1.99
81	<i>Pseudomonas fluorescens</i> PfO-1	6438	60.52	0.003	0.028	0.00443	0.00222		3.18
82	<i>Pseudomonas putida</i> F1	5959	61.86	0.602	0.013	0.00113	0.00187		36.81
83	<i>Ralstonia eutropha</i> H16	2912	66.78	0.238	0.47	0.00483	0.00023	β -Proteobacteria	NA
84	<i>Ralstonia solanacearum</i> GMI1000 chromosome	3716	67.04	0.056	< 0.0001	0.00636	0.00581		22.40
85	<i>Rhizobium etli</i> CFN 42	4381	61.27	0.107	< 0.0001	0.00175	0.01177	α -Proteobacteria	17.65
86	<i>Rhizobium leguminosarum</i> bv. viciae 3841	5057	61.09	0.001	< 0.0001	0.00363	0.01196		NA

A Statistical study on the nucleotide composition of bacterial chromosomes.

87	<i>Rickettsia bellii</i> RML369-C	1522	31.65	0	< 0.0001	0.00859	0.01514		26.08
88	<i>Rickettsia conorii</i> Malish 7	1268	32.44	0.584	0.052	0.00294	0.00634		16.28
89	<i>Rickettsia rickettsii</i> 'Sheila Smith'	1257	32.47	0.575	0.002	0.00182	0.00767		NA
90	<i>Rickettsia typhi</i> Wilmington	1111	28.92	0.919	0.007	0.00020	0.01395		26.15
91	<i>Salmonella enterica</i> subsp. enterica serovar Typhi CT18	4809	52.09	0.267	0.043	0.00151	0.00152	γ-Proteobacteria	9.85
92	<i>Salmonella</i> <i>typhimurium</i> LT2	4857	52.22	0.89	0.585	0.00043	0.00008		3.58
93	<i>Shigella boydii</i> Sb227	4519	51.21	0.571	0.001	0.00022	0.00249		11.05
94	<i>Shigella flexneri</i> 5 8401	4574	50.92	0.48	0.268	0.00147	0.00214		NA
95	<i>Staphylococcus</i> <i>aureus</i> RF122	2742	32.78	0.788	0.427	0.00130	0.00247	Firmicutes	0.10
96	<i>Staphylococcus</i> <i>epidermidis</i> ATCC 12228	2499	32.1	< 0.0001	< 0.0001	0.01246	0.01087		21.12
97	<i>Staphylococcus</i> <i>haemolyticus</i> JCSC1435	2685	32.79	0.001	0	0.00584	0.00643		NA
98	<i>Streptococcus</i> <i>mutans</i> UA159	2030	36.83	0.111	0.046	0.00403	0.00679		
99	<i>Streptococcus</i> <i>pyogenes</i> MGAS2096	1860	38.73	0.619	0.15	0.00133	0.00154		3.71
100	<i>Streptococcus</i> <i>thermophilus</i> CNRZ1066	1796	39.08	0.05	0.863	0.00537	0.00459		2.63
101	<i>Streptomyces</i> <i>coelicolor</i> A3(2)	8667	72.12	0.001	0.037	0.00394	0.00134	Actinobacteria	NA
102	<i>Thermotoga</i>	1860	46.25	0.171	< 0.0001	0.00344	0.01548	Thermotogae	39.15

A Statistical study on the nucleotide composition of bacterial chromosomes.

	<i>maritima</i> MSB8									
103	<i>Thermotoga petrophila</i> RKU-1	1824	46.09	0.733	< 0.0001	0.00013	0.01687		NA	
104	<i>Thiobacillus denitrificans</i> ATCC 25259	2909	66.07	0.962	0.086	0.00027	0.00059	β-Proteobacteria	5.70	
105	<i>Vibrio cholerae</i> O395	3024	47.78	< 0.0001	0.069	0.00514	0.00105		NA	
106	<i>Vibrio fischeri</i> ES114	1332	37.03	<i>0.001</i>	0.037	0.00994	0.00491			
107	<i>Xanthomonas campestris</i> pv. <i>campestris</i> ATCC 33913	5076	65.07	0.196	0.719	0.00302	0.00038			
108	<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> KACC10331	4941	63.69	0.87	0.499	0.00104	0.00065			γ-Proteobacteria
109	<i>Xylella fastidiosa</i> 9a5c	2679	52.68	< 0.0001	< 0.0001	0.04727	0.05291			62.97
110	<i>Xylella fastidiosa</i> Temecula1	2519	51.78	0.044	0	0.00379	0.01093			6.44
111	<i>Yersinia pestis</i> CO92	4653	47.64	0.649	<i>0.001</i>	0.00090	0.00520			NA
112	<i>Yersinia pseudotuberculosis</i> IP32953	4744	47.61	0.969	<i>0.001</i>	0.00124	0.00496			

Chromosomes of bacteria analyzed in this study. Kolmogorav-Smirnov test (KS) for significance between the frequency distribution of complementary nucleotide values are given as KS (W) between A & T and KS (S) between G & C. In bacteria, p -values $< 10^{-4}$ (strong violation of ISFDP) are shown in bold letter and p -values < 0.01 but $\geq 10^{-4}$ (weak violation of ISFDP) are shown in italics. The p -value between 10^{-4} and 10^{-3} is shown as 0.000. Relative absolute abundance value difference between the complementary nucleotides are given by $|(\sum A - \sum T)| / (\sum A + \sum T)$ and $|(\sum G - \sum C)| / (\sum G + \sum C)$ for AT-skew (ATS) and GC-skew (GCS), respectively. In chromosome of *X. fastidiosa* 9a5c the GCS/ATS value is highest suggesting the difference between the abundance values of complementary nucleotides is high. The p

value by KS test is in concordant with the ATS/GCS suggesting that the abundance difference can be represented by the frequency distribution study of the nucleotides. Similar relation is also observed in other chromosomes.

Table 2.2: Summary of the frequency distribution parity test

Organism	Number of chromosomes	Number of chromosomes exhibiting ISFDP for both W & S	Number of chromosomes violating* ISFDP for both W & S	Number of chromosomes violating ISFDP only between S nucleotides	Number of chromosomes violating ISFDP only between W nucleotides
Bacteria	112	60	15 ($5^a + 8^b + 0^c + 2^d$)	30 ($13^e + 17^f$)	07 ($1^g + 6^h$)

*Violation of ISFDP includes both weak ($10^{-2} > P \geq 10^{-4}$) and strong ($P \leq 10^{-4}$).

a Strong violation between S nucleotides as well as between W nucleotides.

b Strong violation between S nucleotides but weak violation between W nucleotides.

c Weak violation between S nucleotides but strong violation between W nucleotides.

d Weak violation between S nucleotides as well as between W nucleotides.

e Strong violation only between S nucleotides.

f Weak violation only between S nucleotides

g Strong violation only between W nucleotides.

h Weak violation only between W nucleotides

Figure 2.1 (a-e): Frequency distribution of nucleotides in chromosomes

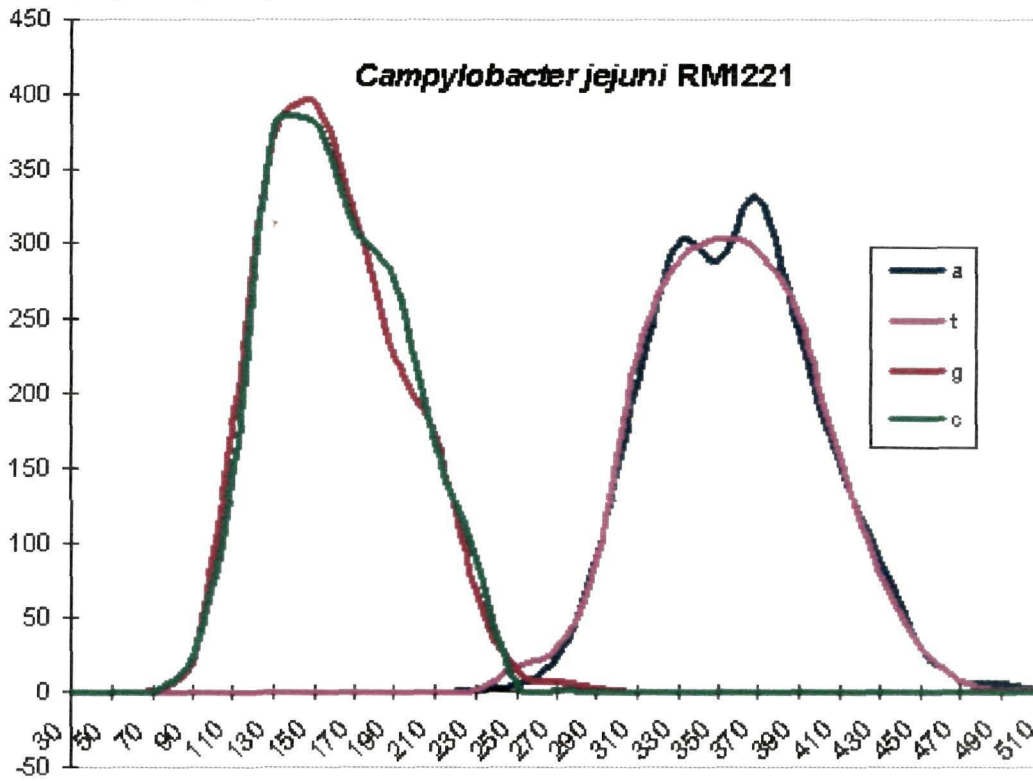


Fig 2.1a(i)

Cumulative distributions (a / t)

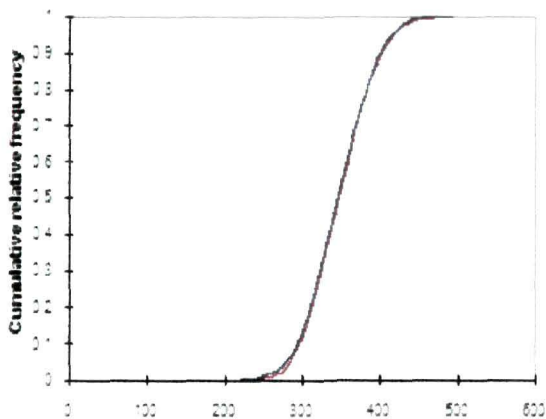


Fig 2.1a(ii)

Cumulative distributions (g / c)

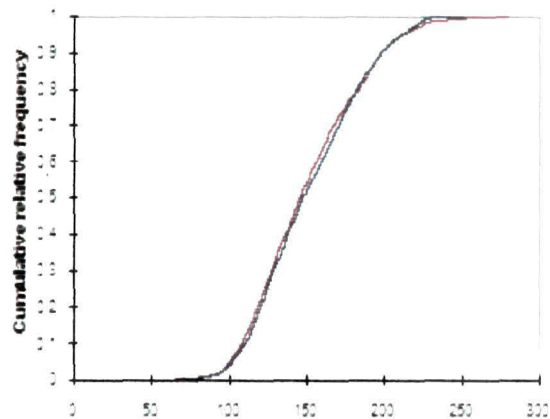


Fig 2.1a(iii)

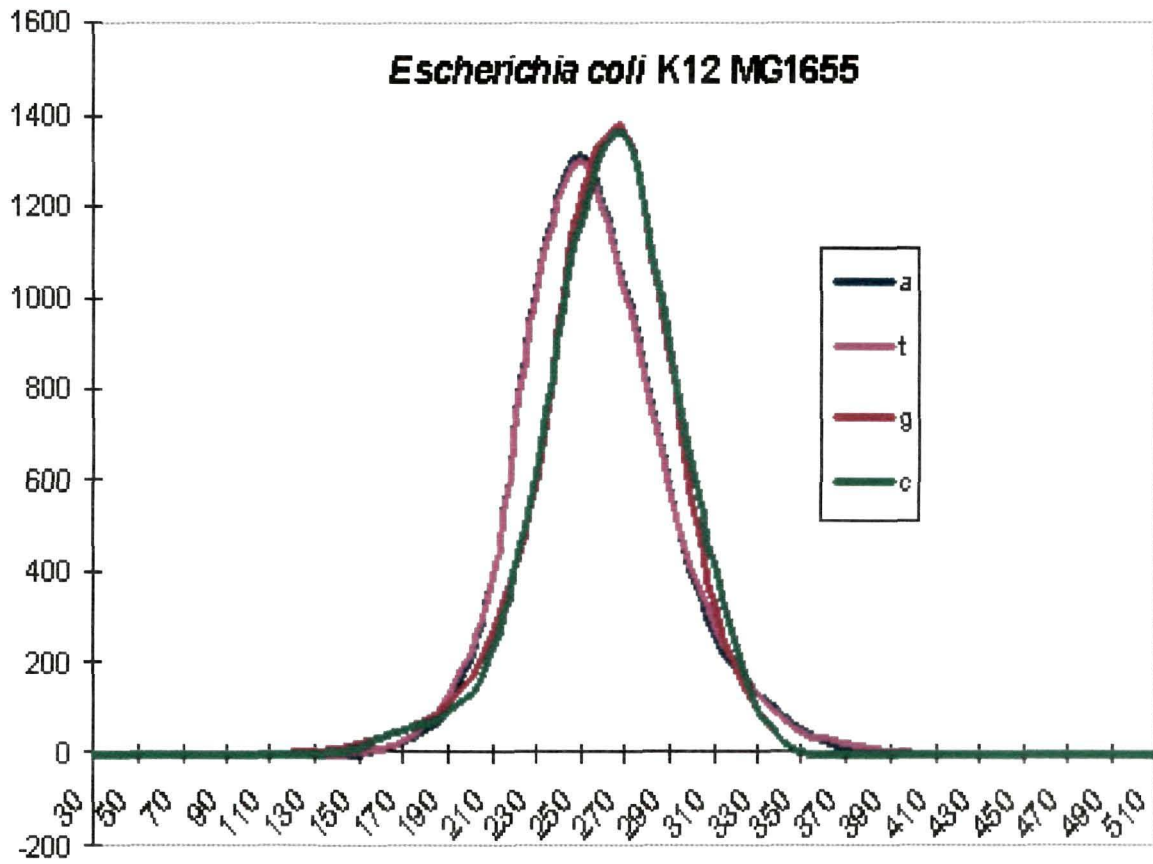


Fig 2.1b(i)

Cumulative distributions (a / t)

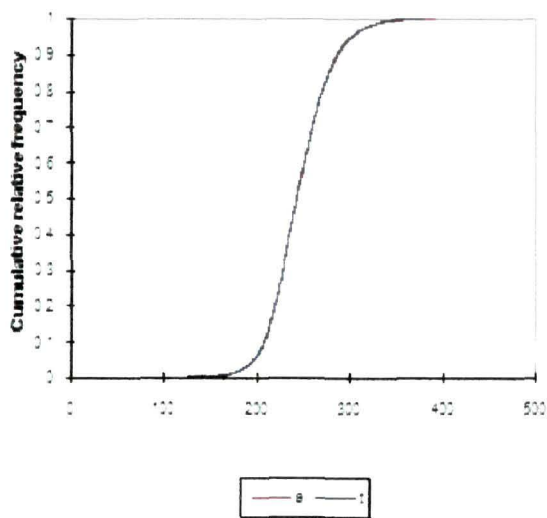


Fig2.1b(ii)

Cumulative distributions (g / c)

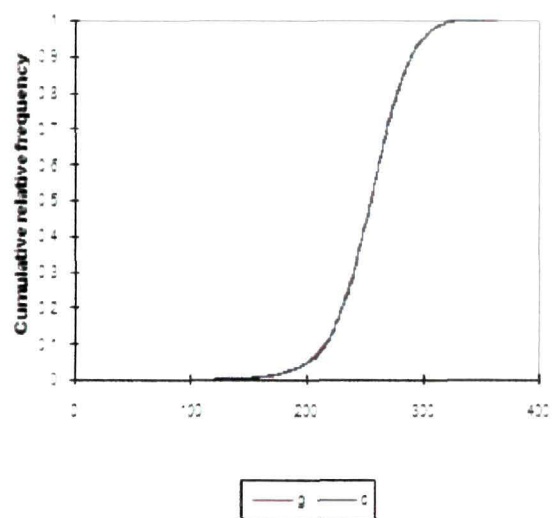


Fig 2.1b(iii)

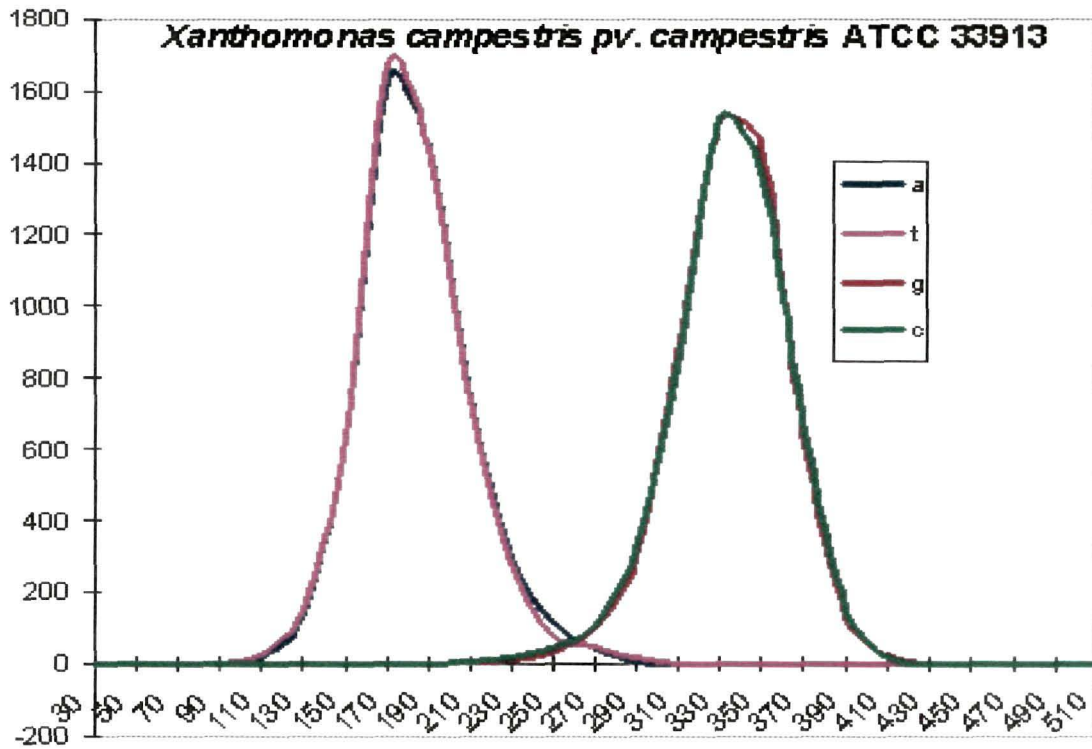


Fig 2.1c(i)

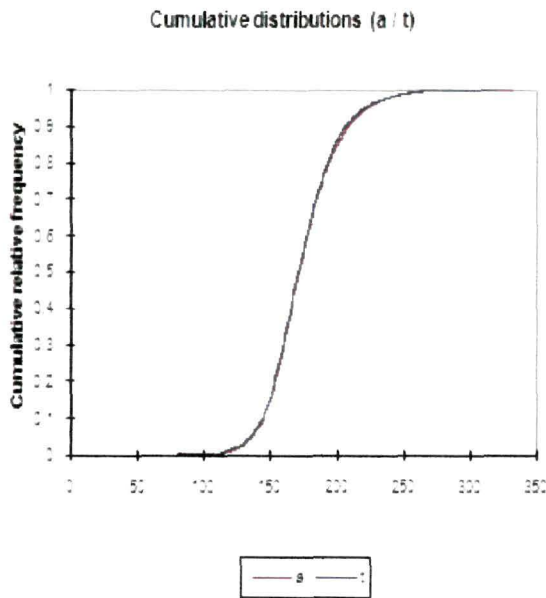


Fig 2.1c(ii)

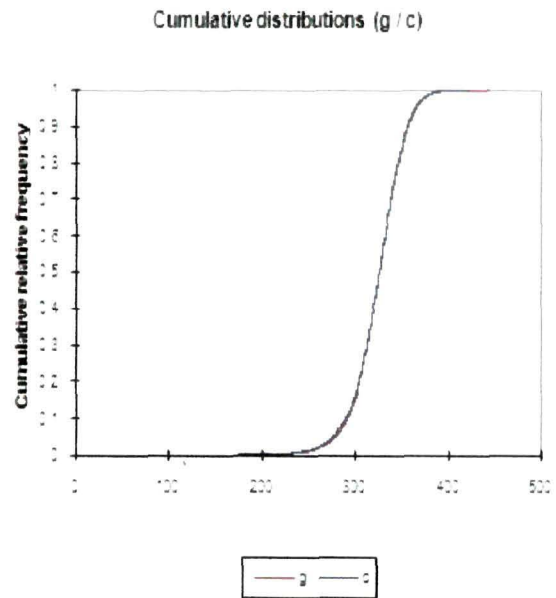


Fig 2.1c(iii)

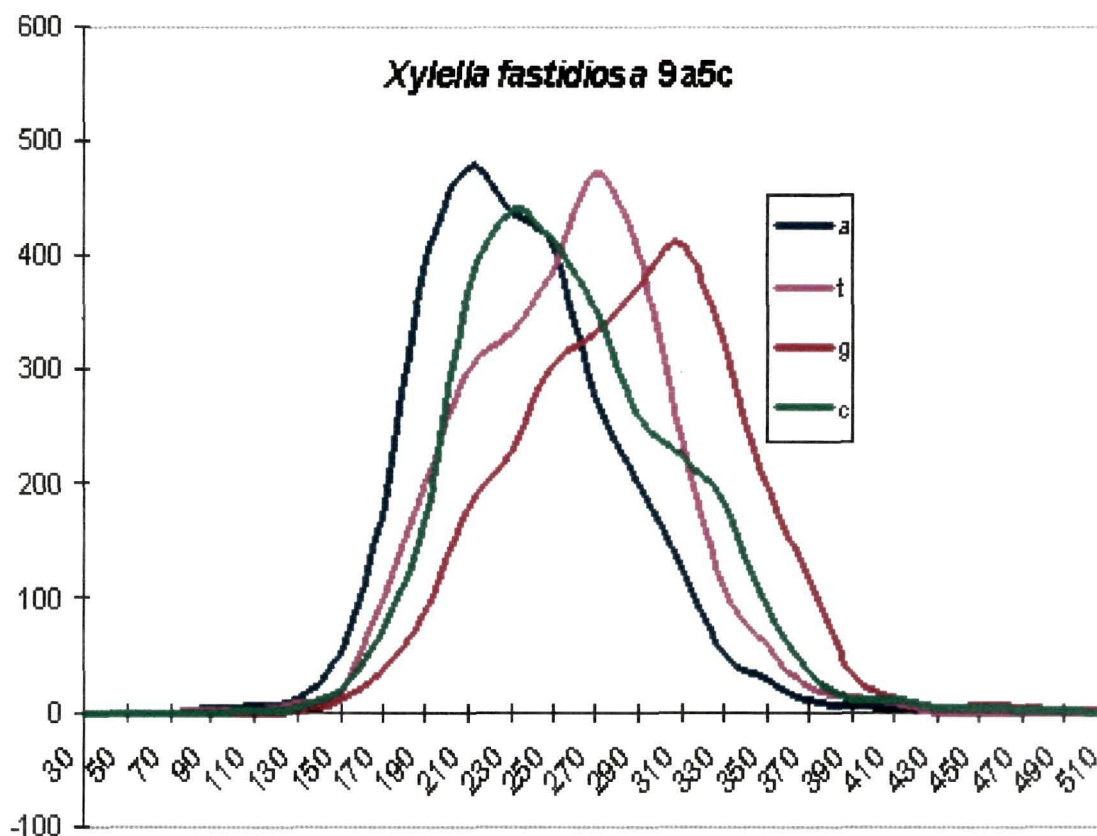


Fig 2.1d(i)

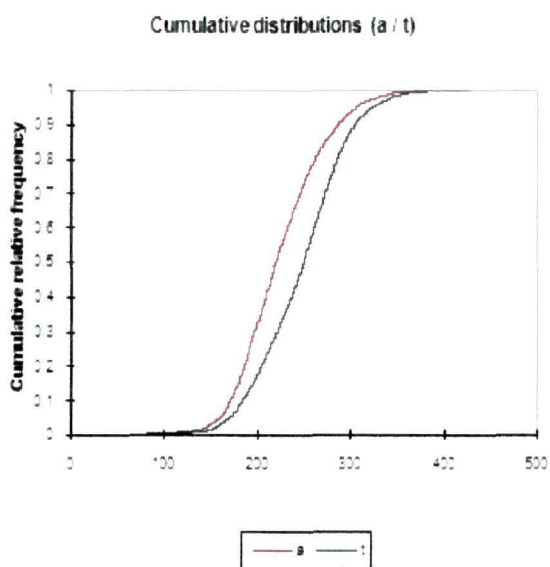


Fig 2.1d(ii)

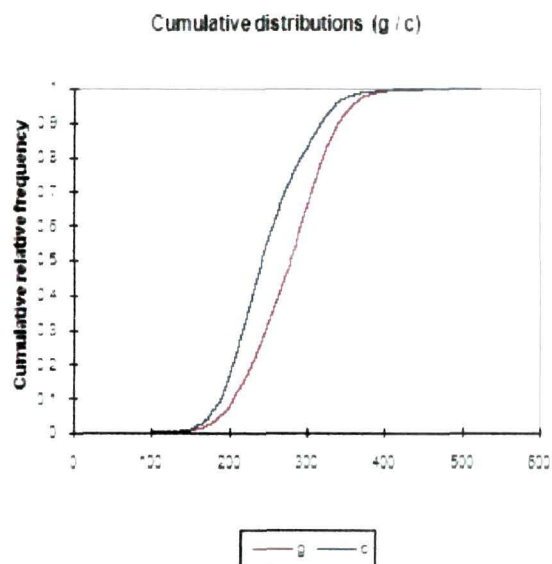


Fig 2.1d(iii)

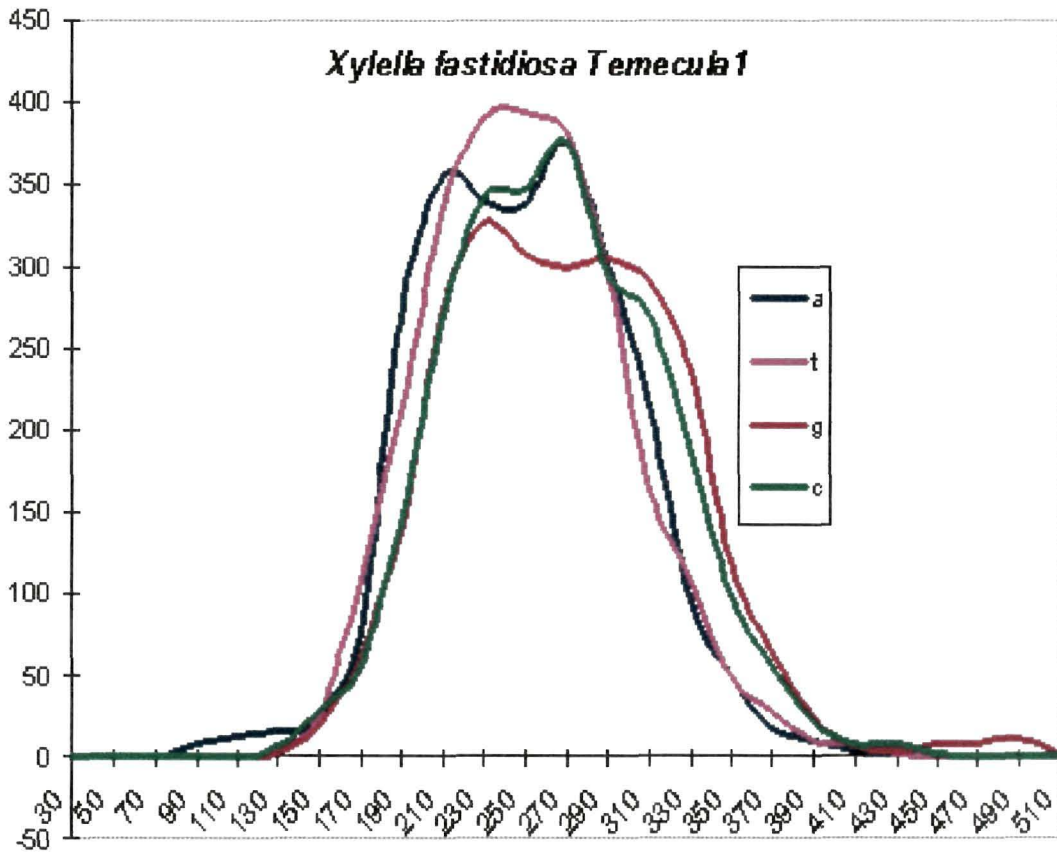


Fig 2.1e(i)

Cumulative distributions (a / t)

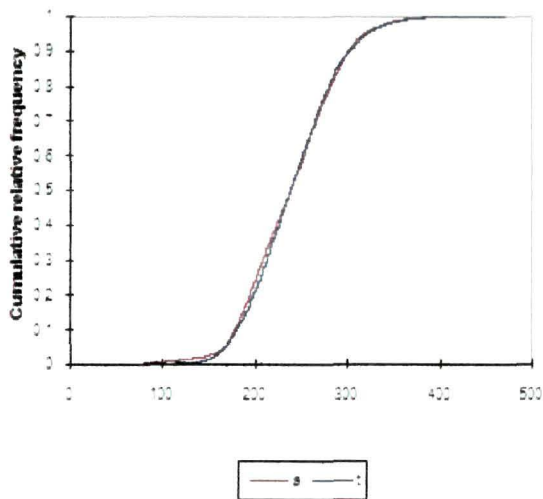


Fig 2.1e(ii)

Cumulative distributions (g / c)

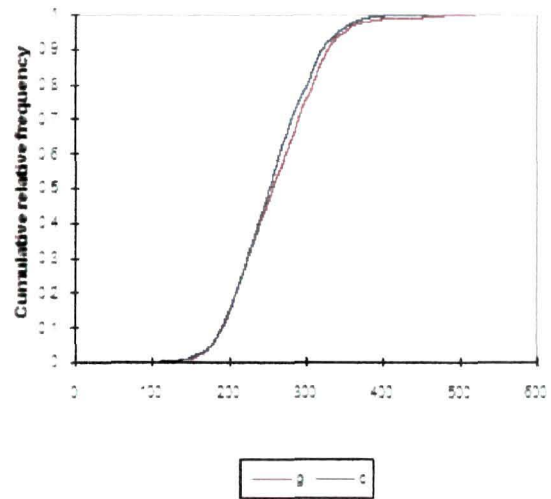


Fig 2.1e(iii)

Fig 2.1(a-e)

Smooth curves present the group frequency distribution of the four nucleotides, a (blue), t (pink), g (red), c (green). X-axis represents the abundance values of the nucleotide spanning a range while the Y-axis represents the frequency of the abundance values. In Fig 2.1a the chromosome is AT rich, in Fig 2.1b the chromosome is composed of similar AT and GC while in Fig. 2.1c the chromosome is GC rich. This is also evident from the group frequency distribution curve. The smooth frequency curves of complementary nucleotides in these chromosomes are overlapping with each other. Kolmogorov-Smirnov test (KS-test) is shown for S and W nucleotides separately adjacent to the figures, respectively [2.1a(ii, iii)---2.1e(ii, iii)]. The KS-test is in concordance with the curve obtained by smoothing group frequency distribution. In Fig. 2.1d and 2.1e the group frequency distribution for the chromosomes of two strains of *Xyllela fastidiosa* is shown. In 9a5c strain chromosome, the smooth frequency curve between the complementary nucleotides does not overlap which is also suggested by the KS-test. However, in Temecula 1 strain chromosome the parity is maintained.

Out of 112 bacterial chromosomes, 60 chromosomes exhibited ISFDP, 15 chromosomes exhibited violation between S as well as between W nucleotides, 30 chromosomes exhibited violation only between S nucleotides and 7 chromosomes exhibited violation only between W nucleotides (Table 2.2). Chromosomes of *Alkaliphilus oremlandii* OhILAs (36.26%), *Agrobacterium tumefaciens* C58 (circular; 59.38%), *Mycobacterium ulcerans* Agy99 (65.47%), *Staphylococcus epidermidis* ATCC 12228 (32.1%), and *Xylella fastidiosa* 9a5c (52.68%) exhibited strong violations between S nucleotides as well as between W nucleotides. Chromosomes of the three *Bacillus anthracis* (35.35%) strains, *Lactobacillus reuteri* F275 (38.87%), *Magnetococcus* sp. MC-1 (54.17%), *Mycobacterium leprae* TN (57.8%), *Rhizobium leguminosarum* bv. *viciae* 3841 (61.09%), and *Rickettsia bellii* RML369-C (31.65%) exhibited strong violation between S nucleotides as well as weak violation between W nucleotides. Chromosomes of *Coxiella burnetii* Dugway 7E9-12 (42.44%) and *Staphylococcus haemolyticus* JCSC1435 (32.79%) exhibited weak violation between S as well as between W nucleotides. Chromosome of *Vibrio cholerae* O395 (47.78%) exhibited strong violation of ISFDP only between W nucleotides. Similarly, there are six chromosomes where weak violations only between W nucleotides were observed. Chromosomes of *Bacillus thuringiensis* serovar *konkukian* 97-27 (34.41%), *Bordetella parapertussis* 12822 (68.1%), *Bordetella pertussis* Tohama 1 (67.72%), *Haemophilus influenzae* PittGG (38.01%), *Helicobacter hepaticus* ATCC 51449 (35.93%), *Lactobacillus acidophilus* NCFM (34.72%), *Lactobacillus brevis* ATCC 367 (46.22%), *Nitrobacter winogradskyi* Nb-255 (62.05%), *Ralstonia solanacearum* GMI1000 chromosome (67.04%), *Rhizobium etli* CFN 42 (61.27%), *Thermotoga maritima* MSB8 (46.25%), *Thermotoga petrophila* RKU-1 (46.09%), exhibited strong violation only between S nucleotides. Similarly there are 17 chromosomes exhibited weak violation only between S nucleotides. The interesting findings came from this study is that violations of ISFDP within a chromosome with respect to S and W nucleotides may not be of similar magnitudes. This study suggests that ISFDP is commonly observed among chromosomes and its violation is not as rare as described earlier (Deng, 2007). ISFDP violation found in bacteria belongs to different groups, possessing different GC% and with different genome sizes.

Usually different strains within a species are found to be similar with respect to ISFDP such as the eight *E. coli* strains were observed to exhibit ISFDP between S nucleotides as well as between W nucleotides, the three *Bacillus anthracis* strains are found to similar in terms of their ISFDP violation (strong violation of ISFDP between S nucleotides as well as weak violations of ISFDP between W nucleotides). However, variation among the strains of a bacterial species with respect to ISFDP was observed as follows: out of the two strains of *Coxiella burnetii*, Dugway 7E9-12 strain violated ISFDP whereas RSA 493 strain exhibited ISFDP. Out of the four *Haemophilus influenza* strains, 86-028NP and PittEE exhibited violation of ISFDP, whereas PittGG and Rd KW20 exhibited strong and weak violations only between S nucleotides, respectively. *Xylella fastidiosa* 9a5c exhibited strong violation of ISFDP whereas *X. fastidiosa* Temecula1 exhibited weak violation of ISFDP only between S nucleotides. These are called as intra species ISFDP violations. Chromosomes of four species of *Mycobacterium* genus exhibited large difference among each other with respect to ISFDP. Chromosome of *Mycobacterium* sp. KMS (68.44%) exhibited parity between S as well as between W nucleotides whereas chromosome of *Mycobacterium ulcerans* Agy99 (65.47%) exhibited strong violation of the parity between S as well as between W nucleotides.

2.4.2. ISFDP weakly correlates with Chargaff's 2nd parity

Comparison of ISFDP was done with the ATS / GCS in chromosomes to find out whether one can define the other. GCS was compared with ISFDP violation between S nucleotides and ATS was compared with ISFDP violation between W nucleotides. Among the bacterial chromosomes, maximum GCS was found in *X. fastidiosa* 9a5c with the value 0.0529. All of the 16 chromosomes with $GCS \geq 0.01$ were found to violate ISFDP (14 strongly violated and 2 weakly violated). Out of the 18 chromosomes with $GCS \geq 0.005$ but < 0.01 , 6 exhibited insignificant violation, 7 exhibited strong violation and 5 exhibited weak violation of ISFDP. Similarly, out of 56 chromosomes with $GCS \geq 0.001$ but < 0.005 , 5 exhibited strong violation, 11 exhibited weak violation and 40 exhibited insignificant violation. Out of the 22 chromosomes with $GCS < 0.001$ except *Bacillus thuringiensis* Al Hakam chromosome (with GCS value 0.00081 exhibited weak violation of ISFDP) other exhibited insignificant violation. Maximum ATS was found in *X. fastidiosa* 9a5c with the value 0.04727. Out of the 5 chromosomes with $ATS \geq 0.01$, 4 were found to violate ISFDP (2

strongly violated and 2 weakly violated) whereas *Mycoplasma hyopneumoniae* J exhibited insignificant violation (with ATS 0.0102). Out of the 14 chromosomes with $ATS \geq 0.005$ but < 0.01 , 6 exhibited insignificant violation; 3 exhibited strong violation and 5 exhibited weak violation of ISFDP. Out of the 67 chromosomes with $ATS \geq 0.001$ but < 0.005 , 57 exhibited parity, 1 strongly violated and 9 violated weakly between the W nucleotides. All the 26 chromosomes with $ATS \leq 0.001$ exhibited insignificant violation of ISFDP. These results suggest that chromosomes with high ATS/GCS (≥ 0.01) have a stronger propensity to violate ISFDP and chromosomes with low ATS/GCS (≤ 0.001), have a stronger propensity to exhibit ISFDP. However, chromosomes with intermediate ATS/GCS (≥ 0.001 and ≤ 0.01) have the possibility of either exhibiting parity or violating the parity.

Correlation analysis was done between the p-values (from KS-test between) of W nucleotides and ATS as well as between the p-values (from KS-test between) of S nucleotides and GCS. The r-values are -0.5572 and -0.4526 for W and S nucleotides, respectively. This suggests that the correlation between the two intra-strand parity features is weak. The correlation between ATS and GCS is 0.629, which suggests that parity violation between S nucleotides weakly correlates with parity violation between W nucleotides within a chromosome. Unlike ATS and GCS correlation, no correlation was found between p-values (KS-test) of W and p-values (KS-test) of S nucleotides, which supports that ISFDP and Chargaff's 2nd parity are not identical.

2.4.3. The chromosomes with asymmetric replication topography are more prone to ISFDP violation in bacteria

Bacterial chromosome is a single replicon. Due to bi-directional mode of replication, one part of a strand is synthesized as LeS where as the other part is synthesized as LaS. In most of the chromosomes, the mutational strand asymmetry causes K nucleotides $>$ M nucleotides in LeS and the reverse in (K nucleotides $<$ M nucleotides) in LaS. In an ideal case where the termination site is located symmetrically with respect to the origin of replication in a chromosome, the excess of K nucleotides in LeS will be similar to the excess of M nucleotides in LaS and therefore will cancel each other to exhibit Chargaff's 2nd parity in chromosomes. Potential replication origin and termination sites for different chromosomes based on AT skew (ATS), GC skew (GCS), coding sequence skew (CDS), nucleotide skew at

the 3rd position of codons and oligonucleotides skew in chromosomes have been reported (Frank and Lobry, 2000; Worning *et al.*, 2006), which has been reviewed in detail (Sernova and Gelfand, 2008). Out of the 112 bacterial chromosomes analyzed in this study, information regarding potential site for the origin and termination of 57 chromosomes is available. ISFDP violation between S nucleotides was compared with angular deviation of termination site because G > C in LeS is a more universal feature of chromosomes than T > A in LeS. Of the 112 chromosomes, maximum angular deviation of 71.28° is reported in *Bordetella pertussis* Tohama 1. Out of the 14 chromosomes where $\geq 20^\circ$ angular deviation was observed, 12 exhibited violation of ISFDP between S nucleotides. *Pseudomonas putida* F1 (61.86%) with 36.8° and *Coxiella burnetii* RSA 493 (42.66%) with 31.14° angular deviations exhibited insignificant parity violation. Out of the 11 chromosomes with deviation $\geq 10^\circ$ but $< 20^\circ$, 4 chromosomes exhibited ISFDP violation between S nucleotides. Out of the 30 strains with deviation $\geq 1.0^\circ$ and $\leq 10^\circ$, 9 chromosomes exhibited parity violation between S nucleotides. *Chlamydomonas abortus* S263 with angular deviation only 0.569°, parity violation was observed only between S nucleotides. This study indicates that chromosomes with higher asymmetric topography are more prone to violate the parity. However, chromosomes with symmetric replication topography were also observed to violate the parity.

The correlation coefficient between angular deviations and GCS as well as ATS values are 0.474 and 0.357, respectively, suggesting a weak correlation. The correlation between angular deviations and p-value of S (KS-test between S nucleotides) as well as p-value of W (KS-test between W nucleotides) are -0.259 and -0.048, respectively. The angular deviation in *X. fastidiosa* 9a5c is 62.96° whereas the same in Temecula 1 is 6.44°. The difference in the magnitude of ISFDP violation between the strains might be attributed to the chromosome topography. Comparison for the four *H. influenzae* strains could not be done due to the unavailability of information for all the strains. The Rd KW20 chromosome (that violated ISFDP) has the angular deviation 46° might be an important factor to violate ISFDP.

2.4.4. Composition of forward encoded and reverse encoded sequences within DNA strands might influence the parity

Most of the regions in prokaryotic chromosomes are composed of coding sequences. Presence of both forward encoded and reverse encoded sequences in bacterial chromosomes

has been proposed for the observation of Chargaff's 2nd parity in chromosomes (Baisnée *et al.*, 2002, Verma *et al.* 2005) So we analyzed only coding sequences in chromosomes of bacteria to study ISFDP as follows (Fig. 2.2): In one way (Case I), a DNA strand is only composed of only forward encoded sequences and in the other way (Case II) a DNA strand is composed of 50% forward encoded and 50% reverse encoded sequences. The result is shown for *E. coli* chromosome [Fig. 2.3(a&b)]. The smooth frequency curves of complementary nucleotides overlap in Fig. 2.3 (b) whereas in Fig. 2.3(a) they do not overlap. The significance of these overlaps were studied by KS-test which suggests that the similarity between the distribution of complementary nucleotides in Case II. Similar results were obtained by analysis of several (ten) other bacterial chromosomes.

A comparative analysis between the Ws and Cs in a chromosome with respect their composition of forward encoded sequences was done in *X. fastidiosa* species as well as in *H. influenza* species. The relative difference of the compositional abundance values forward sequences in Ws and Cs of *X. fastidiosa* 9a5c and *X. fastidiosa* Temecula 1 chromosomes are 0.078 and 0.015, respectively, which indicates that the proportion of forward encoded and reverse encoded sequence in 9a5c strain is more disproportionate than that of Temecula 1 strain, which might be the reason for a stronger parity violation in the former. The relative difference of the compositional abundance values forward encoded sequences in Ws and Cs of *H. influenzae* 86-028NP (exhibits parity) and *H. influenzae* Rd KW20 (violates parity) chromosomes are 0.030 and 0.005, respectively, which suggests that the proportion of forward encoded and reverse encoded sequence in 86-028NP strain is more disproportionate than that of Rd KW20 strain. This is in contrast to the result is of *X. fastidiosa*, i.e. parity violation is observed in the strain (Rd KW20) with more proportionate gene distribution between Ws and Cs whereas insignificant parity violation is observed in chromosome with disproportionate gene distribution between the strands. A quantitative estimation of the coding sequences in both the strands of the chromosomes was done in few other bacteria such as *A. tumefaciens*, *B. subtilis*, and *E. coli* (Fig. 2.4). *A. tumefaciens* was shown to possess minimum relative difference of ORF numbers between the strands but violates parity whereas *E. coli* and *B. subtilis* having greater gene composition bias between the strands exhibits parity. The results from this indicate that a higher disproportionate composition of forward

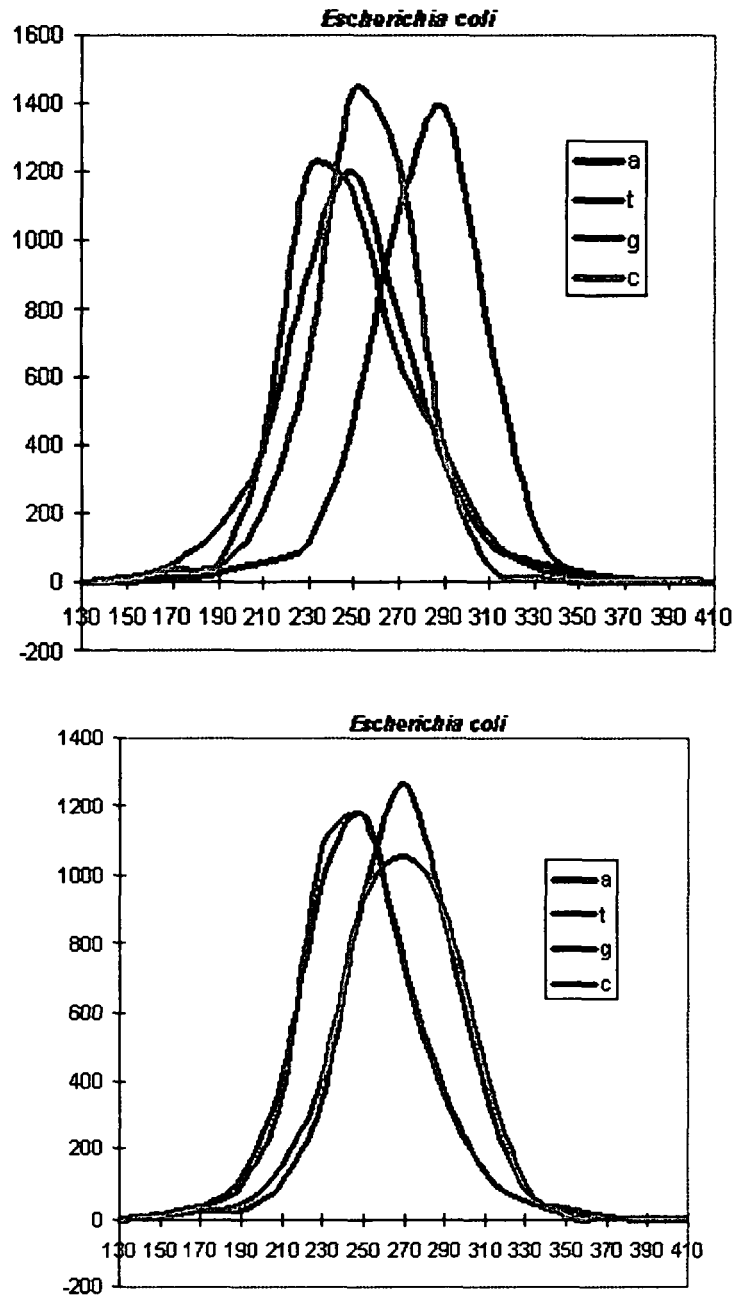
and reverse encoded sequences within a strand has greater propensity to parity violation. However, proportionate composition of the sequences not necessarily implies the exhibition of parity.

Figure 2.2: Schematic representation of coding sequence arrangement studied



In the upper case the entire DNA strand is composed of forward encoded sequences (black arrows). Parity is not observed in this case. In the lower case the DNA strand is made up of 50% forward encoded sequences and the other 50% is the reverse encoded sequences (white colored). Parity is observed in this case.

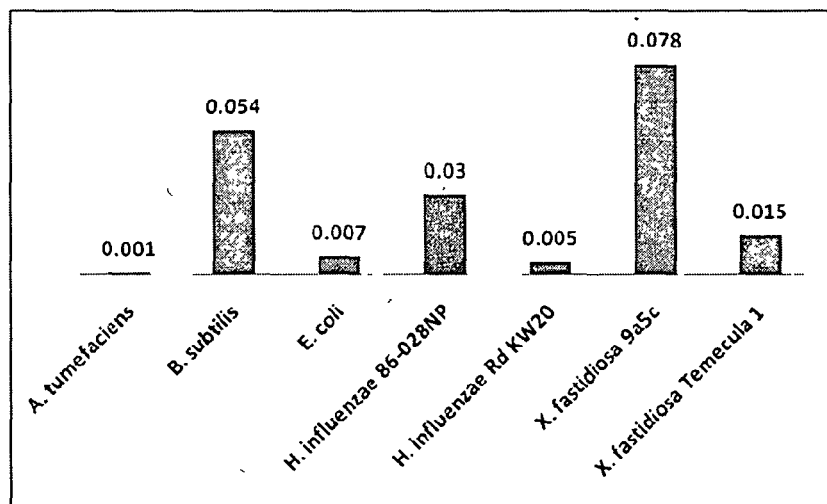
Figure 2.3 (a, b) : Frequency distribution study of nucleotides in coding sequences

**Fig. 2.3 (a-b)**

Smooth curves present the group frequency distribution of the four nucleotides a (blue), t (pink), g (red), c (green). X-axis represents the abundance values of the nucleotide spanning a range while the Y-axis represents the frequency of the abundance values. In Fig. 2.3a the frequency of the nucleotides in a DNA strand only composed of forward encoded sequences

of *E. coli* is shown (coding sequences analyzed for other chromosomes exhibited the similar feature). It is evident from the Fig. 2.3a that the complementary nucleotides frequency distributions of the complementary nucleotides do not overlap. In Fig 2.3b, the frequency of the nucleotides of the same DNA strand done where 50% of the sequence was joined with the rest after reverse complementation (Materials and Methods). This resembled a strand composed of 50% forward encoded sequences and 50% reverse encoded sequences. It is evident from the figures that parity between the complementary nucleotides is observed in this case. These observations have been confirmed by KS-test.

Figure 2.4: Relative disproportionate composition of ORFs between Ws and Cs in Chromosomes



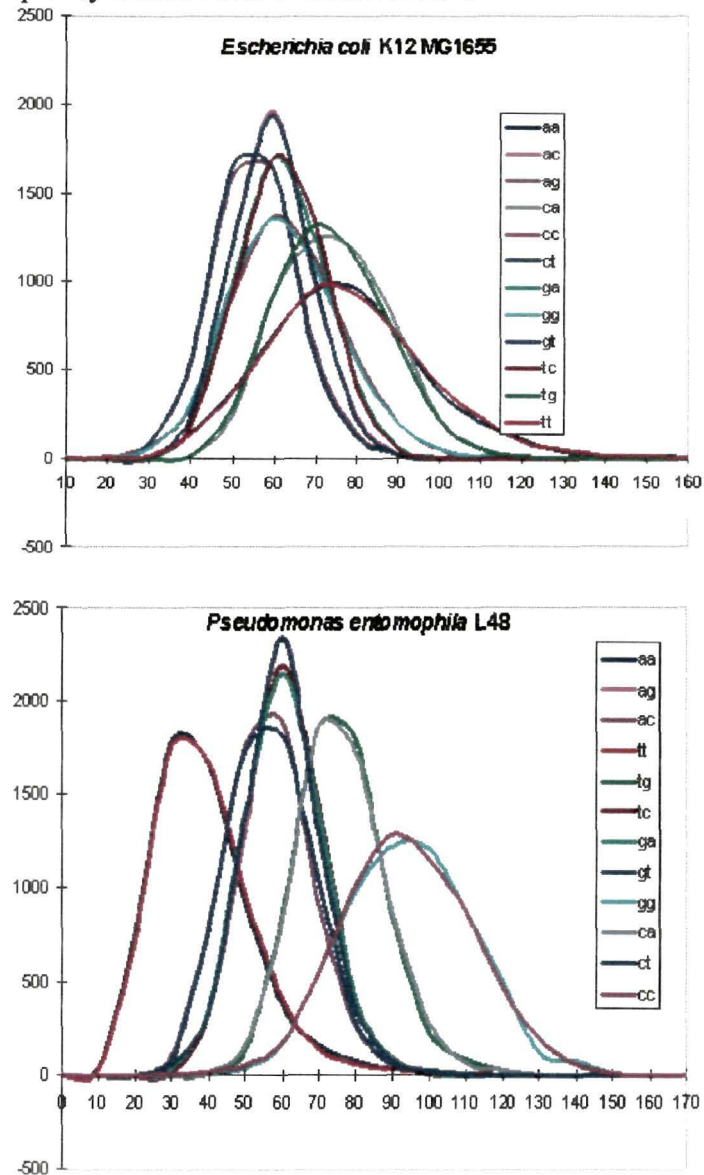
The composition of ORFs in Ws and Cs of seven bacteria was studied. Relative disproportionate composition was found out by deducting the ORF numbers between the two strands and then dividing the value obtained with the total number of ORFs present in both strands. X-axis represents the bacteria while the Y-axis represents the relative disproportionate. In *A. tumefaciens* relative disproportionate value found to be minimum suggesting that the difference in the number of ORFs between the strands is relatively

minimum in comparison to others. Both *A. tumefaciens* exhibited ISFDP violations whereas insignificant ISFDP violation observed between *E. coli*, *B. subtilis*. Comparison between strains of *X. fastidiosa* as well as *H. influenzae* is shown.

2.4.5. Intra-strand frequency distribution parity between complementary oligonucleotides in chromosomes

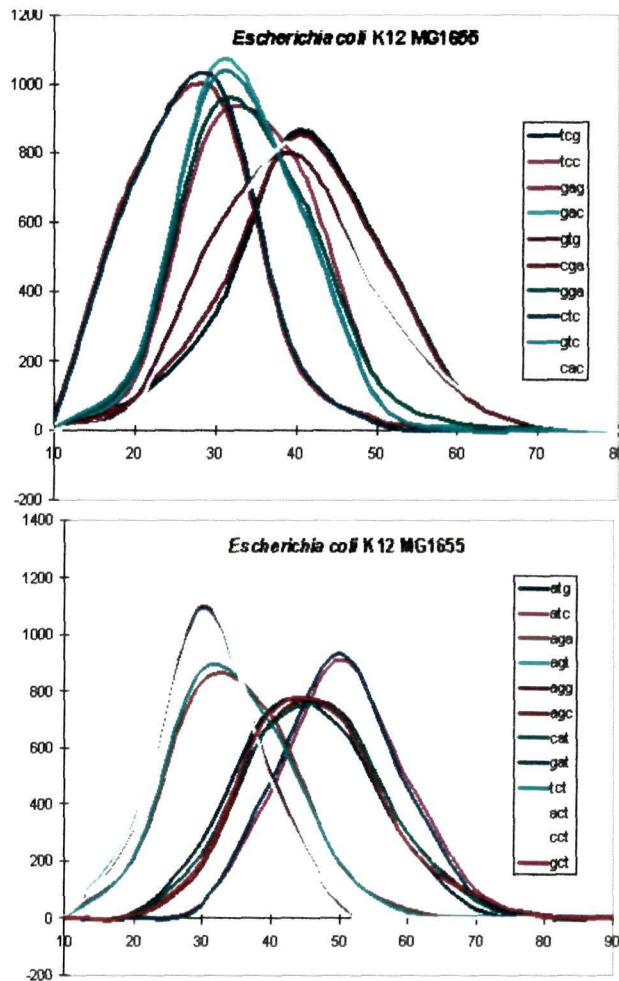
Intra-strand parity between compositional abundance values of complimentary oligonucleotides is well reported. We studied here frequency distribution of complementary di- and tri-nucleotides in chromosomes as described for mononucleotides. The smooth curves of oligonucleotide frequencies have been shown in extra figures. In Fig. 2.5 (a & b), the frequency distributions of dinucleotides has been shown for *E. coli* K12 MG1655 and *Pseudomonas entomophila* L48 chromosome (64.16%). Out of the 12 smooth frequency curves (four palindromic dinucleotides were excluded), overlapping of the curves between complementary dinucleotides is observed. In Fig 2.5a, though the abundance values of aa, tt, tg and ca dinucleotides in *E. coli* chromosome are close, the distributions between the complementary dinucleotides are found only overlapping and that of the non-complementary ones are different. The distributions for aa and tt follow a higher standard deviation (values not shown) than that of tg and ca. Similarly, gg and cc dinucleotides distributions exhibit a higher standard deviation (values not shown) than that of the dinucleotides tc and ga, though the abundance values of the four dinucleotides are close to each other. The significance of the similarity was studied by KS test which suggested that frequency distributions between complementary dinucleotides are statistically similar. Apart from this, di-nucleotides distribution parity has been studied in three more bacterial chromosomes and similar result has been observed. In Fig. 2.6 (a & b) distribution of twenty two trinucleotides of *E. coli* K12 MG1655 chromosome is shown. Like dinucleotides, overlapping between the distributions of complementary tri-nucleotides is also observed. Distribution similarity between complementary trinucleotides was studied by KS-test for the 64 trinucleotides which suggested that distributions of complementary trinucleotides within a strand are similar. The same study was done in one more bacterial chromosome (data not shown) and similar results were obtained.

Figure 2.5(a, b): Frequency distribution of dinucleotides



Smooth curves present the group frequency distribution of the twelve non-palindromic dinucleotides in the chromosomes *E. coli* and *P. entomophila* are shown. X-axis represents the abundance values of the dinucleotides spanning a range while the Y-axis represents the frequency of the abundance values. The frequency distribution between complementary dinucleotides (Fig. 2.5 a & b) is so well that the twelve curves can be grouped under six curves. KS-test suggests similarity between the distributions of the complementary dinucleotides. The distributions of two non-complementary dinucleotides are found to be different by KS-test.

Figure. 2.6(a, b): Frequency distribution of trinucleotides in Escherichia coli chromosome

**Fig. 2.6(a,b)**

Smooth curves present the group frequency distribution of the twenty two trinucleotides in the chromosome of *E. coli*. X-axis represents the abundance values of the trinucleotides spanning a range while the Y-axis represents the frequency of the abundance values. It is evident from the figures that the frequency distributions of the complementary trinucleotides are similar. This has been studied for the 64 trinucleotides and similar result was observed for all.

2.5. Discussion

We have described in this study a new intra-strand parity feature in chromosomes, which is found in bacteria. This parity is also found in archaea and eukaryotic chromosomes (Powdel *et al.*, 2009). The methodology used to study this parity gives the statistical significance of similarity between the two distributions of complementary nucleotides/oligonucleotides. The basic qualitative feature of ISFDP is not changing for a chromosome even the segmentation is done at random taking any point out of the first 1000 nucleotides as the starting point. In other words sampling fluctuation is not affecting the feature. The correlation between the ISFDP and ISP is not strong, which is in accordance with the view that similarity in the total abundance values of two complementary nucleotides will not always yield similarity in their frequency distribution pattern. However, violation of ISP will definitely exhibit violation of ISFDP. Around 50% of the chromosomes in bacteria are found to exhibit ISFDP violations. Chromosomes of *H. influenzae* Rd KW20, *M. tuberculosis* F11, etc which has been reported to exhibit ISP are found to violate ISFDP (Shioiri and Takahata, 2001).

ISFDP violation observed in all possible combinations in chromosomes: i. the violation of parity between S nucleotides as well as between W nucleotides; ii. only between S nucleotides; and only between W nucleotides. The correlation between ATS and GCS is found to be not strong suggesting that parity violation between S nucleotides not necessarily always associate with parity violations between W nucleotides and the *vice versa*. This can be called as intra-chromosomal parity violations. ISFDP violations of different magnitudes were found among chromosomes of different strains belonging to a species which can be referred as intra-species parity violations. Examples are *Coxiella burnetii*, *H. influenzae* and *X. fastidiosa*. These intra-chromosomal and intra-species violations suggest that there may not be any strict rule existing in cells to maintain ISFDP in chromosomes. Differential ISP among chromosomes within a species and between chromosomes within a bacterium has already been reported in *Chlamydomyphila pneumoniae* strains and *Deinococcus radiodurans* R1 chromosomes (Shioiri and Takahata, 2001), respectively. However, these were not considered significant in their study due to lack of statistical proof. Oligonucleotide skew patterns also have been found to be variable among strains of *Yersinia pestis*. These intra-

species variation of chromosomal features is interesting and needs in-depth analysis of the genome sequences to find out the reason which might reveal the reason for ISP/ISFDP violation in chromosomes and between the two intra-strand parity features.

Enrichment of LeS with K nucleotides over M nucleotides and the *vice versa* in LaS due to the strand mutational asymmetry is a general observation in chromosomes. Due to bidirectional replication, GCS/ATS in LeS is cancelled with GCS/ATS in LaS which results into the establishment of parity in chromosomes. The cancellation effect indirectly suggests that the compositional abundance values between two complementary nucleotides even though they differ within a sub-chromosomal region. This is in support of the observation here that chromosomes with higher GCS/ATS values are violating the ISFDP and chromosomes with lower GCS/ATS are exhibiting the parity. However, the chromosomes with intermediate range GCS/ATS are found to exhibit parity as well as violate parity and this violation is independent of genome GC%. For example *Streptococcus mutans* UA159, *Rickettsia conorii* Malish 7, *Campylobacter jejuni* subsp. *jejuni* 81116, *Campylobacter concisus* 13826 and *Lactococcus lactis* subsp. *cremoris* MG1363, *Helicobacter pylori* J99 are (all AT rich organisms) chromosomes with $GCS \geq 0.005$, exhibits ISFDP between S nucleotides whereas chromosomes of *Bacillus anthracis* strains (AT rich) with similar GCS (> 0.005), violate ISFDP between S nucleotides. So ISFDP in these chromosomes is an interesting aspect of future research.

In concordant with the view of the bidirectional replication and establishment of parity in chromosomes, several chromosomes with higher asymmetric replication topography were found to violate the ISFDP. The exceptions are *Pseudomonas putida* F1, and *Coxiella burnetii* RSA 493 chromosomes with 36° and 31° angular deviations, respectively. Chromosomes of *Chlamydomphila abortus* S263 and *Magnetospirillum magneticum* AMB-1, with very less angular deviations 0.57° and 2.14° , respectively are found violating ISFDP. This indicates that features apart from the replication topography might contribute to the parity establishment in chromosomes. Proportionate composition of forward encoded sequences between the two strands though thought to be responsible to establish the parity after analysis of the artificially constructed chromosomes, several observations went against it. The extreme case is *A. tumefaciens* where the composition is very much proportionate but

violations of ISFDP are strong. So the two factors such as asymmetric replication topography and disproportionate composition of forward encoded sequences between the strands in chromosomes that were assumed to play important roles in determining ISFDP violations were found to be insufficient.

In spite of different selection/mutation pressures on chromosomes as exemplified by codon usage (Sharp *et al.*, 2005), replication topography (Frank and Lobry, 2000), isochores (Duret *et al.*, 2006), GCS/ATS (Grigoriev, 1998) the tendency of the chromosomes of all types towards maintaining the ISFDP is interesting. Since ISFDP and ISP are outcomes of compositional abundance of nucleotides (mono/oligo), theories proposed for ISP might hold true for ISFDP. The Nussinov-Forsdyke hypothesis is that stem-loop potential has an adaptive advantage and therefore an important factor driving the compositional symmetry (ISP) between complementary oligonucleotides (Nussinov, 1984; Forsdyke, 1995) has been challenged recently by Chen and Zhao (2005) for human chromosomes. This indicates that the stem-loop (recombination) hypothesis might not be the only explanation for ISP in chromosomes. Baisnée *et al.*, (2002) have argued that the reverse complement symmetry does not result only from point mutation or from recombination, but from a combination effect of different mechanisms at different orders (Baisnée *et al.*, 2002). Two independent reports have theoretically shown that multiple inversion events in chromosomes can establish ISP (Albrecht-Buehler, 2006; Okamura *et al.*, 2007). Though this hypothesis looks fine theoretically, frequent inversion unable to explain the universal observation of opposite GCS/ATS in LeS and LaS (Rocha, 2004), gene distribution asymmetry between the strands (Rocha and Danchin, 2003) and maintenance of gene orders among different bacterial chromosomes (Rocha, 2008). This hypothesis also does not describe any functional significance/advantage of ISP/ISFDP feature, which is so wide spread in chromosomes. Theoretically, it has also been argued that mismatch error repairing system is responsible to establish Chargaff's second parity rule in chromosomes (Deng, 2007). However, the intra-chromosomal parity violation observed in this study goes against this hypothesis.

We think the important factor that determines ISP/ISFDP in chromosomes is the bidirectional replication. This causes one part of a strand Ws/Cs as LeS and the other part as LaS. The strand mutational asymmetry and gene distribution asymmetry between LeS and

LaS therefore cancel out each other within the strand to exhibit the parity. In case of ssDNA/ssRNA viruses, gene distribution is restricted to one strand only depending on which these are called as either + or – strand viruses. The genome size is also not large (< 10 kb) in these phages (Adams and Antoniw, 2005, 2006) and during replication, one strand only acts as the template on which the other strand is made. Most likely these features are responsible for violating the parity in these genomes. The advantage of bidirectional replication in bacteria and archaea where the nucleus is absent, are as follows: i. quicker completion of replication than the unidirectional mode of replication and ii. the meeting of the two replication forks might be sending some signal to the cell for the completion of chromosome replication where the nucleus is absent. Symmetric replication topography will help to terminate the replication from the origin in a lesser time in comparison to an asymmetric topography. So the selection pressure to maintain symmetric replication topography in fast growing bacteria is likely to be more than in slow growing bacteria. This proposition has similarity with the Selection Mutation Drift theory proposed for codon usage (Bulmer, 1991) in bacteria. Our study of ISFDP of *Vibrio* species (the generation time is 0.2 to 0.3 hour; fast growing) in this context seems to be also not holding true here because its chromosomes violate ISFDP between W nucleotides. Moreover, comparison of generation time (Rocha, 2004) with asymmetry in replication topography of chromosomes (Worning *et al.*, 2006) exhibits no correlation (data not shown). More research on this aspect will give conclusive result if growth rate has any relation with parity establishment in chromosomes. In conclusion our study has revealed an interesting aspect of intra-strand parity. Future research will reveal the reason for the presence of this parity in chromosomes.

CHAPTER III

3. Strand-specific mutational bias influences codon usage of weakly expressed genes in *Escherichia coli*

3.1. Abstract

According to the selection-mutation-drift (SMD) theory of molecular evolution, mutation predominates in determining codon usage bias in weakly expressed genes while selection predominates in determining codon usage bias in highly expressed genes. Strand-specific mutational bias causes compositional asymmetry of the nucleotides between leading and lagging strands in bacterial chromosomes. Keeping in view the above points, codon usage bias between the strands were compared in *Escherichia coli* chromosome. In comparison to highly expressed genes, codon usage of weakly expressed genes was observed to be more biased towards strands: G ending codons were significantly more in leading strands than lagging strands and the reverse was true for the C ending codons. In case of weakly expressed genes, the GC₃ skews were found to be significantly different between the strands. This suggests that strand-specific mutational bias influences codon usage of weakly expressed genes to a greater extent than that of highly expressed genes. The differential effect of strand-specific mutational bias in *E. coli* might be attributed to stronger purifying selection in the highly expressed genes than the weakly expressed genes. The observation here in *E. coli* supports the selection-mutation-drift theory of molecular evolution.

3.2. Introduction

Nonrandom usage of synonymous codons, otherwise called as codon usage bias (CUB), is common in prokaryotes and eukaryotes. Patterns and degrees of CUB vary not only among different organisms, but also among genes in the same genome (Grantham *et al.*, 1980a; Ikemura, 1985). CUB is affected by both mutation and selection pressures in organisms (Bulmer, 1991, Osawa *et al.*, 1992; Hershberg and Petrov, 2008). Among the mutation pressures, genome G+C content (GC%) is important (Sueoka, 1962; Muto and Osawa, 1987). GC% ranges from 17 (*Candidatus Carsonella ruddii* PV; NC_008512) to 75 (*Anaeromyxobacter dehalogenans* 2CP-C; NC_007760) among bacteria. In these genomes a greater variation of GC%, ranging from <10 to >90, occurs at the third position of codons (GC₃), the most neutral position. Most of the nuclear genome of warm-blooded vertebrates is

a mosaic of very long (>>200 kilobases) DNA segments, known as isochores, distinguished by differences in their GC% (Bernardi *et al.*, 1985). CUB is directly linked to variability of GC% among isochores that affect both coding and introns or intergenic regions (Se'mon *et al.*, 2006). Therefore, GC% variation is the most important parameter influencing CUB among different organisms (Chen *et al.*, 2004). Apart from the GC%, other biases (Rocha, 2004) that affect the relative frequency of synonymous codons include strand specific mutational bias (Lobry, 1996) and the transcription coupled repair associated bias (Francino and Ochman, 2001). Different evidences supporting the role of selection in CUB are as follows: (i). CUB of highly expressed genes is different from that of weakly expressed genes (Gouy and Gautier, 1982; Duret and Mouchiroud, 1999). CUB and gene expression are positively correlated (Sharp and Li, 1986a, 1987b; dos Reis *et al.*, 2003); (ii). Positive correlation between the abundance of tRNAs in the cytosol and the occurrence of the respective codons in genes (Ikemura, 1981; Ikemura, 1985; Duret, 2000). Codons corresponding to abundant tRNAs are more frequently present in the highly expressed genes for efficient translation; (iii). The rate of synonymous substitution between species is inversely proportional to the CUB in the genes (Sharp and Li, 1987a).

Intra-genomic codon usage variation between highly expressed genes and weakly expressed genes is explained by two theories: (i) The expression-regulation theory, which states that both highly expressed genes and weakly expressed genes are equally under purifying selection for the presence of optimal and non-optimal codons, respectively, to keep their expression high or weak (Grosjean and Fiers 1982; Konigsberg and Godson, 1983; Walker *et al.*, 1984, Hinds and Blake 1985); (ii) The selection-mutation-drift theory (SMD), which states that mutation predominates in determining CUB in weakly expressed genes while selection predominates in determining CUB in highly expressed genes (Sharp and Li, 1986a, 1986b; Bulmer 1987, 1988, 1991). Though SMD theory is widely accepted now, certain questions are yet to be answered. The inter-genomic sequence divergence in case of weakly expressed genes was observed to be more in comparison to that of highly expressed genes among enterobacteria such as *Escherichia coli*, *Salmonella typhimurium*, *Klebsiella pneumoniae* etc (Sharp and Li, 1987a). This was explained by the higher intensity of selection at the translational level in highly expressed genes than in weakly expressed genes

(Sharp and Li, 1987a): However, a later study in enterobacteria, by comparing the substitution rate in codon families such as lysine and phenylalanine, suggested that the decrease in the mutation rate in highly expressed genes in comparison to weakly expressed genes is responsible for the decline substitution (Eyre-Walker and Bulmer, 1995). Since transcription coupled repair acts less frequently on the template DNA of weakly expressed genes than that of highly expressed genes, the former is more prone to mutation than the latter (Eyre-Walker and Bulmer, 1995). So the role of selection or mutation for the variability among genes in rates of synonymous substitution is still debatable. Similarly, evidence in support of a mutational bias associated with replication occurring more in weakly expressed genes than in highly expressed genes within a genome is yet to be demonstrated. Our endeavour here is to address the later problem by studying the effect of strand-specific mutational bias on the two gene types in *E. coli* chromosome.

The strand-specific mutational bias causes compositional asymmetry between the leading strands (LeS) and the lagging strands (LaS) in bacteria. In most of the bacteria higher frequency of the keto nucleotides (G & T) is observed in the LeS in comparison to the LaS (Frank and Lobry, 1999; Lobry and Sueoka, 2002). In some other cases such as Firmicutes (gram-positive bacteria with low GC%) with heterodimeric DNA polymerase III α -subunit constituted by PolC and DnaE, higher frequency of the purine nucleotides (G & A) is observed in the LeS in comparison to the LaS (Freeman *et al.*, 1998; Hu *et al.*, 2007). The similar compositional biases observed between the strands are caused by very different mutational effects among organisms (Rocha *et al.*, 2006). The best known example of strand-specific mutational bias affecting CUB is *Borrelia burgdorferi*. In this bacterium, CUB in genes is affected by the mode of replication i.e. either LeS or LaS and not by the expression (McInerney, 1998). It is not known whether the influence of strand-specific mutational bias on CUB in genes is independent of their expression. *B. burgdorferi* is not an ideal case to address this question because the strength of selected codon usage bias 'S' has been reported to be low in this bacterium (Sharp *et al.*, 2005). The strength of selected codon usage bias has been reported to be variable among bacteria: the species exposed to selection for rapid growth have more rRNA operons, more tRNA genes and more strength of selected codon usage bias (Sharp *et al.*, 2005).

In case of *E. coli* the strength of selected codon usage bias is high (Sharp *et al.*, 2005). The proteome has been well quantified in this bacterium thereby giving the information about the expression levels of different genes (Ishihama *et al.*, 2008). The empirically determined expression values strongly correlate with the theoretically determined codon adaptation index (Sharp and Li, 1987b; dos Reis *et al.*, 2003; Ishihama *et al.*, 2008). In addition, the bacterium has been well characterized at growth (Bremer and Dennis, 1996) as well as genome (Touchon *et al.*, 2009) levels. Therefore, this is an ideal case to address the question regarding the differential influence of strand-specific mutational bias on the two gene types.

We compared codon usage between LeS and LaS of *E. coli* chromosome. Our hypothesis in this study is that strand-specific mutational bias influences CUB in weakly expressed genes to a greater extent in comparison to that of highly expressed genes because of stronger purifying selection in the latter than the former. The observation in this study for *E. coli* goes in favor of this hypothesis. So, this study is in favour of the SMD theory of molecular evolution.

3.3. Materials and Methods

3.3.1. Separation of highly expressed genes and weakly expressed genes in LeS and LaS

Coding sequences of *E. coli* K12 MG1655 chromosome were downloaded from the DDBJ website (<http://gib.genes.nig.jp/>). Highly expressed genes (HEG) and weakly expressed genes (WEG) were separated by using the information of cytosolic protein abundance values (Ishihama *et al.*, 2008) and codon adaptation index (CAI; Sharp and LI, 1987b). All the genes selected as HEG and WEG were of length more than 100 amino acids. We took 100 genes from the top with maximum expression as HEG, 100 genes from the bottom with minimum expression as WEG. CodonW (<http://www.molbiol.ox.ac.uk/cu/tutorial.html>; Peden, 1999) was used to calculate codon adaptation index [CAI; Sharp and Li, 1987b, GC3 (GC% at the 3rd position)] and sizes of genes. Accordingly, genes with expression levels $3.69 < \log_2 x < 7.10$ [$\log_2 x$ is the logarithm of protein copy number (x) per cell to the base 2] were considered as HEG and genes with

expression levels $1.8 < \log_2 x < 2.25$ were considered as WEG (Table 3.3). Genes with low expression were in concordant with their low CAI (Table 3.3).

The replication origin and terminus points for the separation of LeS and LaS were taken from the website (Worning *et al.*, 2006; <http://www.cbs.dtu.dk/services>).

3.3.2. Codon usage study between LeS and LaS

Synonymous codon frequency (SCF) of family box codons is similar to relative synonymous codon usage (RSCU) described by Sharp *et al.* (Sharp *et al.*, 1986). SCF is defined as

$$SCF_{ACA}^F = \frac{X_{ACA}}{\sum_{N \in \{A, T, G, C\}} X_{ACN}}$$

Superscript 'F' stands for family box.

We used SCF instead of RSCU because only family box codons were considered here. X_{ACA} is defined as the abundance values of ACA in a group of genes considered under HEG (or WEG) taken from either LeS or LaS. $\sum X_{ACN}$ is defined as the summation of the abundance values of the four synonymous codons in family box ACN of the genetic code.

Absolute difference of SCF between LeS and LaS of a codon was calculated to find its usage difference between the strands. The statistical significance of the difference was tested using Z-test.

Change in relative synonymous codon usage (CRSCU) was measured to find the differential usage of family box codons between the strands as follows:

$$CRSCU(aa) = \sum_{xyz \in \mathcal{A}} |SCF_{LeS}(xyz) - SCF_{LaS}(xyz)|$$

CRSCU (aa) stands for relative synonymous codon usage for the family box codons of an amino acid 'aa'. xyz represents a codon. $SCF_{LeS}(xyz)$ and $SCF_{LaS}(xyz)$ are the respective values of SCF(xyz) in LeS and LaS, respectively. In this study CRSCU has been calculated for eight family box codons of the eight amino acids.

For the purpose of group wise comparison of the strand specific mutational bias in the genes in ascending order of expression we modified the formula to account only the mutational biases arising in G and C bases in the LeS and LaS. We revised the formula as follows-

$$\text{CRSCU (GC)} = \sum_{xyz \in A, z \in (G, C)} |SCF_{LeS}(xyz) - SCF_{LaS}(xyz)|$$

To measure the replication bias and transcription-translation induced bias in a group of genes of similar expressions we used the measures B_I and B_{II} defined by Lobry and Sueoka (2002), which are given by

$$B_I = \sqrt{\left\{ \left(\frac{G_3}{G_3 + C_3} \right)_{LeS} - \left(\frac{G_3}{G_3 + C_3} \right)_{LaS} \right\}^2 + \left\{ \left(\frac{A_3}{A_3 + T_3} \right)_{LeS} - \left(\frac{A_3}{A_3 + T_3} \right)_{LaS} \right\}^2}$$

Where $\left(\frac{G_3}{G_3 + C_3} \right)_{LeS}$ is the average value of $\left(\frac{G_3}{G_3 + C_3} \right)$ for all the genes in the LeS and similar definitions for other three ratios. Now taking

$$x_1 = \frac{1}{2} \left\{ \left(\frac{G_3}{G_3 + C_3} \right)_{LeS} + \left(\frac{G_3}{G_3 + C_3} \right)_{LaS} \right\} \quad \text{and}$$

$$y_1 = \frac{1}{2} \left\{ \left(\frac{A_3}{A_3 + T_3} \right)_{LeS} + \left(\frac{A_3}{A_3 + T_3} \right)_{LaS} \right\}$$

B_{II} is defined as

$$B_{II} = \sqrt{(0.5 - x_1)^2 + (0.5 - y_1)^2}$$

3.3.3. ATS_3 and GCS_3 between LeS and LaS

ATS_3 and GCS_3 are the third position AT and GC skews respectively. These are calculated as $ATS_3 = [A_3/(A_3+T_3)]$ and $GCS_3 = [G_3/(G_3+C_3)]$ (Sueoka, 1999). The difference

of the skews in LeS and LaS for a family box was compared within individual groups (WEG and HEG). The significance of the difference was tested using Z- test.

3.3.4. Estimation of strand-specific mutational bias

The strand-specific mutational bias with respect to the abundance values of G & C nucleotides as well as A & T nucleotides was estimated in highly expressed genes, weakly expressed genes and intergenic regions. In case of the coding region only abundance values at the 3rd position of codons were considered. ΔGC : is the average difference of GC skews $[(G-C)/(G+C)]$ between the strands (LeS - LaS). ΔAT : is the average difference of AT skews $[(A-T)/(A+T)]$ between the strands (LeS - LaS).

3.3.5. Transfer RNA gene ratio

Transfer RNA gene ratio per amino acid was calculated by dividing the total number of isoacceptor tRNA genes with the total number of synonymous codons for that amino acid (Satapathy *et al.*, 2010). Information relating to tRNA gene numbers was taken from Genomic tRNA Database (<http://gtrnadb.ucsc.edu>), which uses tRNAscan-SE to classify tRNA into different groups by studying their anticodon sequence (Lowe and Eddy, 1997).

3.4. Results

3.4.1. Study of synonymous codon usage bias in ascending order of gene expression

The eight hundred ninety three *E. coli* genes with known expressions (Ishihama *et al.*, 2008), were arranged in ascending order of their protein abundance values. The genes were divided into total nine groups from the lowest expression level (group 1) towards the highest expression level (group 9). Though correlation was observed between CAI values and expression of the genes, this correlation was not observed in individual groups. This is because genes with similar expression were found with variable CAI values. Number of genes in each group varies in the range 94 to 101 (Table 3.1). Gene compositional asymmetry between the strands (LeS and LaS) was observed in all the groups with the maximum in group 9 having the most highly expressed genes. We studied in each group the effect of

strand specific mutational bias at the 3rd position of codons using three measures: strand specific mutational bias due to replication (B_I), transcription-translation associated bias (B_{II}), and change in relative synonymous codon usage (CRSCU) (Table 3.1). B_I and B_{II} measures were defined by Lobry and Sueoka (2002). CRSCU was measured only for G and C ending codons because strand specific mutational bias in *E. coli* is already known to occur with respect to these nucleotides (Lobry, 1996). Pearson's correlation coefficients of the three measures mentioned above with average expression of the genes in each group was found to be significant (Table 3.1): (i). the correlation coefficient between B_I and average gene expression is -0.708 ($p < 0.01$).

The -ve correlation suggests that the influence of strand specific mutational bias due to replication is more in weakly expressed genes than in highly expressed genes; (ii). the correlation coefficient between B_{II} and average gene expression is 0.963 ($p < 0.0001$). This indicates that transcription-translation associated bias goes on increasing with the expression level of genes; (iii). the correlation coefficient between CRSCU and average gene expression is -0.812 ($p < 0.001$). This indicates that flexibility of codon usage is more towards the weakly expressed gene than the highly expressed genes in *E. coli*. Similar result was obtained when the genes were divided into five different groups instead of nine by combining the two adjacent groups into one starting from the group 1 (Table 3.2).

Table 3.1: Correlation in the expression rank order and the CRSCU

Groups	No. of genes	Gene Distribution asymmetry (LeS/LaS)	Average expression (\log_2 protein abundance)	B_I	B_{II}	CRSCU (GC)
Gr-1	100	1.500	2.08	0.06709	0.15263	0.59263
Gr-2	100	1.564	2.29	0.04180	0.14264	0.52693
Gr-3	99	1.250	2.44	0.04928	0.17100	0.55925

Gr-4	100	1.439	2.59	0.04450	0.16016	0.50949
Gr-5	100	1.564	2.74	0.02894	0.19067	0.37438
Gr-6	100	1.128	2.89	0.05786	0.17988	0.53026
Gr-7	101	1.590	3.09	0.05809	0.21853	0.57327
Gr-8	99	2.000	3.46	0.04647	0.24986	0.43854
Gr-9	94	4.529	4.46	0.00384	0.28679	0.22892
r^*				-0.70832	0.96339	-0.81206

Gr-1 denotes the group of weakly expressed genes and Gr-5 is the group of highly expressed genes

Where B_I is the replication induced biases and B_{II} is the transcription and translation induced biases in a particular group of genes were obtained using the methods shown by Lobry and Sueoka (2002).

** r is the Pearson's correlation coefficient of a particular column with group average gene expression (column 4). All the r -values are significant ($p < 0.01$)*

Table 3.2: Correlation in the expression of rank order (revised group) and the CRSCU

Groups	Gr-1	Gr-2	Gr-3	Gr-4	Gr-5	r(Pearson)
No. of genes	200	199	200	200	94	
Group avg gene Expression	2.18	2.52	2.81	3.27	4.46	
CRSCU(GC)	0.54148	0.50849	0.43388	0.43442	0.22892	-0.98143*

Gr-1 denotes the group of weakly expressed genes and Gr-5 is the group of highly expressed genes

**highly significant*

3.4.2. Strand specific mutational bias between the strands is higher in case of weakly expressed genes than highly expressed genes

We did further a comparative study between hundred most highly expressed and the hundred most weakly genes chosen among the eight hundred ninety three genes (Table 3.3). All the hundred weakly expressed genes considered were with codon adaptation index < 0.4 . Genes having expression level low but having CAI ≥ 0.4 were not considered as weakly expressed genes because these genes might behave as highly expressed genes under some conditions. The range of the protein abundance (\log_2 abundance values) values was 3.69 - 7.10 in case of the highly expressed genes and the same was 1.80 - 2.25 in case of the weakly expressed genes. Out of the 100 highly expressed genes, 81 (23089 codons) were in LeS and 19 were in LaS (6195 codons). Out of the 100 weakly expressed genes, 62 (33281 codons) were in LeS and 38 (23064 codons) were in LaS. Highly expressed genes were relatively smaller than the weakly expressed genes (Ishihama et al., 2008).

Nucleotide frequencies were found out in intergenic regions of *E. coli* chromosome and the frequencies were compared with the nucleotide frequencies at the 3rd position of codons in the two groups of genes (Table 3.4). In the intergenic regions, G & T were observed more in LeS than LaS and the reverse was true in case of A & C nucleotides. In comparison to A & T nucleotides, frequencies of G & C nucleotides were found to be more strand biased, which had already been reported for *E. coli* (Lobry, 1996). The strand biasness of the nucleotide frequencies at the 3rd position of codons were found in the weakly expressed genes. However, the strand biasness of the nucleotide frequencies was found to be lower in case of highly expressed genes. The difference of the skews between the strands was observed in the order as follows: weakly expressed genes $>$ intergenic regions $>$ highly expressed genes (Table 3.4), which is in concordant with the findings reported earlier (Lobry and Sueoka, 2002). The difference between the two types of genes suggests that strand-specific mutational bias is more effective against the weakly expressed genes than highly expressed genes.

Table 3.3: List of highly expressed genes and weakly expressed genes with their strand location of *E. coli* MG1655 chromosome analyzed in this study

Serial No	Gene	CAI	GC ₃	GC	Length*	log ₂ x
Highly expressed genes						
Leading strand						
1	<i>talB</i> __	0.594	0.562	0.522	317	3.913284
2	<i>dnaK</i> __	0.717	0.511	0.512	638	4.494155
3	<i>aceE</i> __	0.668	0.584	0.531	887	4.161368
4	<i>aceF</i> __	0.614	0.551	0.548	630	3.904174
5	<i>rpsB</i> __	0.772	0.548	0.516	241	4.996512
6	<i>tsf</i> __	0.769	0.487	0.501	283	4.819544
7	<i>frr</i> __	0.568	0.486	0.51	185	4.139879
8	<i>tig</i> __	0.732	0.529	0.513	432	4.494155
9	<i>ybaB</i> __	0.64	0.57	0.529	109	3.815578
10	<i>ahpC</i> __	0.797	0.525	0.503	187	5.294466
11	<i>pal</i> __	0.677	0.549	0.516	173	3.704151
12	<i>ompX</i> _	0.736	0.552	0.517	171	4.494155
13	<i>serC</i> _	0.398	0.497	0.503	362	3.758912
14	<i>rpsA</i> _	0.776	0.513	0.513	557	5.421604
15	<i>fabD</i> _	0.456	0.503	0.553	309	4.0086
16	<i>fabG</i> _	0.453	0.474	0.516	244	4.139879
17	<i>icd</i> __	0.564	0.504	0.503	416	3.913284
18	<i>oppA</i> _	0.41	0.487	0.478	543	4.079181
19	<i>yncE</i> _	0.402	0.535	0.508	353	3.913284
20	<i>rplT</i> _	0.675	0.414	0.486	118	4.20412
21	<i>infC</i> _	0.369	0.503	0.489	180	4.421604
22	<i>eda</i> __	0.591	0.565	0.565	213	3.840106
23	<i>yebC</i> _	0.637	0.479	0.523	246	3.913284
24	<i>gnd</i> __	0.537	0.512	0.503	468	4.152288

25	<i>fabB_</i>	0.629	0.602	0.567	406	4.155336
26	<i>guaB_</i>	0.631	0.518	0.547	488	3.93044
27	<i>ndk_</i>	0.638	0.5	0.527	143	3.718502
28	<i>glyA_</i>	0.663	0.57	0.535	417	4.021189
29	<i>rplS_</i>	0.633	0.438	0.487	115	4.287802
30	<i>grpE_</i>	0.495	0.473	0.506	197	4.20412
31	<i>eno_</i>	0.839	0.498	0.505	432	4.746634
32	<i>fbaA_</i>	0.772	0.568	0.513	359	4.679428
33	<i>iktA_</i>	0.686	0.587	0.558	663	3.708421
34	<i>pnp_</i>	0.675	0.538	0.541	711	3.755875
35	<i>rpsI_</i>	0.778	0.476	0.528	130	3.954725
36	<i>rplM_</i>	0.662	0.489	0.509	142	4.78533
37	<i>rplQ_</i>	0.548	0.496	0.541	127	5.075547
38	<i>rpoA_</i>	0.444	0.52	0.524	329	3.895975
39	<i>rpsD_</i>	0.544	0.498	0.51	206	4.959518
40	<i>rpsK_</i>	0.584	0.405	0.519	129	3.913284
41	<i>rpsM_</i>	0.451	0.483	0.523	118	4.494155
42	<i>rplO_</i>	0.695	0.437	0.539	144	4.591065
43	<i>rpsE_</i>	0.58	0.431	0.505	167	4.656098
44	<i>rplR_</i>	0.61	0.431	0.541	117	5.220108
45	<i>rplF_</i>	0.61	0.471	0.524	177	4.269513
46	<i>rpsH_</i>	0.584	0.476	0.505	130	3.985875
47	<i>rpsN_</i>	0.527	0.49	0.531	101	5.946943
48	<i>rplE_</i>	0.602	0.529	0.501	179	4.521138
49	<i>rplX_</i>	0.575	0.485	0.465	104	5.802089
50	<i>rplN_</i>	0.487	0.538	0.515	123	4.320146
51	<i>rplP_</i>	0.599	0.477	0.522	136	4.056905
52	<i>rpsC_</i>	0.727	0.464	0.512	233	4.914343
53	<i>rplV_</i>	0.564	0.505	0.5	110	7.10721
54	<i>rplB_</i>	0.707	0.449	0.529	273	4.651278

A Statistical study on the nucleotide composition of bacterial chromosomes.

55	<i>rplW_</i>	0.651	0.464	0.487	100	5.656098
56	<i>rplD_</i>	0.694	0.526	0.531	201	4.421604
57	<i>rplC_</i>	0.707	0.473	0.517	209	4.901458
58	<i>rpsJ_</i>	0.571	0.604	0.56	103	4.10721
59	<i>tufA_</i>	0.817	0.543	0.534	394	4.951338
60	<i>fusA_</i>	0.744	0.496	0.509	704	4.726727
61	<i>rpsG_</i>	0.531	0.401	0.503	179	5.10721
62	<i>rpsL_</i>	0.658	0.488	0.538	124	4.78533
63	<i>slyD_</i>	0.68	0.536	0.556	196	3.767156
64	<i>asd_</i>	0.359	0.565	0.543	367	3.727541
65	<i>yifE_</i>	0.414	0.422	0.485	112	4.10721
66	<i>udp_</i>	0.536	0.521	0.543	253	3.796574
67	<i>sodA_</i>	0.711	0.602	0.536	206	4.567026
68	<i>rplK_</i>	0.698	0.522	0.526	142	4.0086
69	<i>rplA_</i>	0.768	0.493	0.519	234	4.943495
70	<i>rplJ_</i>	0.631	0.421	0.527	165	4.276462
71	<i>rplL_</i>	0.841	0.265	0.466	121	6.760422
72	<i>rpoB_</i>	0.631	0.578	0.527	1342	3.747412
73	<i>groL_</i>	0.789	0.497	0.53	548	4.78533
74	<i>efp_</i>	0.682	0.47	0.496	188	3.767156
75	<i>hfq_</i>	0.41	0.53	0.493	102	3.767156
76	<i>purA_</i>	0.627	0.538	0.54	432	3.854913
77	<i>rpsF_</i>	0.671	0.528	0.529	131	4.453318
78	<i>rplI_</i>	0.726	0.426	0.501	149	4.20412
79	<i>deoC_</i>	0.63	0.611	0.556	259	4.826723
80	<i>deoA_</i>	0.472	0.58	0.553	440	4
81	<i>deoD_</i>	0.653	0.568	0.526	239	4.660865
Lagging strand						
82	<i>can_</i>	0.39	0.55	0.509	220	3.693727
83	<i>dapD_</i>	0.516	0.571	0.533	274	4.09691

84	<i>gpmA_</i>	0.576	0.529	0.512	250	4.158362
85	<i>pflB_</i>	0.774	0.564	0.513	760	4.082785
86	<i>fabA_</i>	0.533	0.54	0.533	172	4.369216
87	<i>ompA_</i>	0.785	0.546	0.539	346	4.475671
88	<i>hns_</i>	0.582	0.379	0.47	137	5.365488
89	<i>adhE_</i>	0.651	0.49	0.507	891	3.7348
90	<i>fabI_</i>	0.598	0.538	0.534	262	4.09691
91	<i>tpx_</i>	0.544	0.527	0.516	168	4.10721
92	<i>sodB_</i>	0.546	0.508	0.504	193	3.767156
93	<i>gapA_</i>	0.835	0.509	0.502	331	5.230449
94	<i>ackA_</i>	0.656	0.555	0.522	400	4.037426
95	<i>ptsI_</i>	0.496	0.495	0.497	575	3.755875
96	<i>crr_</i>	0.604	0.53	0.475	169	5.482874
97	<i>glnA_</i>	0.622	0.567	0.53	469	3.883661
98	<i>tpiA_</i>	0.742	0.492	0.528	255	4.462398
99	<i>ppa_</i>	0.653	0.588	0.515	176	3.854913
100	<i>yjgF_</i>	0.605	0.563	0.534	128	3.815578
Weakly expressed genes						
Leading strand						
1	<i>djlA_</i>	0.359	0.56	0.528	271	2.222716
2	<i>murE_</i>	0.384	0.563	0.566	495	2.238046
3	<i>murG_</i>	0.313	0.507	0.562	355	2.20412
4	<i>yadG_</i>	0.32	0.535	0.497	308	2.184691
5	<i>yahI_</i>	0.348	0.619	0.563	316	2.167317
6	<i>yaiW_</i>	0.294	0.566	0.535	364	2.222716
7	<i>kefA_</i>	0.365	0.598	0.519	1120	2.143015
8	<i>cusB_</i>	0.35	0.591	0.558	407	2.167317
9	<i>ftsK_</i>	0.361	0.49	0.541	1329	2.214844
10	<i>pyrD_</i>	0.321	0.453	0.471	336	2.238046
11	<i>uup_</i>	0.388	0.56	0.524	635	2.149219

12	<i>helD_</i>	0.303	0.544	0.531	684	1.937518
13	<i>torC_</i>	0.387	0.541	0.514	390	2.167317
14	<i>nagZ_</i>	0.331	0.488	0.532	341	2.222716
15	<i>dada_</i>	0.328	0.579	0.564	432	2.136721
16	<i>sohB_</i>	0.327	0.526	0.507	349	2.149219
17	<i>tyrR_</i>	0.276	0.543	0.522	513	2.161368
18	<i>ydbD_</i>	0.213	0.379	0.435	768	2.08636
19	<i>hrpA_</i>	0.376	0.601	0.535	1300	2.017033
20	<i>pqqL_</i>	0.28	0.441	0.467	931	1.882525
21	<i>ydeP_</i>	0.281	0.472	0.504	759	2.0086
22	<i>hipA_</i>	0.225	0.489	0.481	440	2.068186
23	<i>yneE_</i>	0.263	0.498	0.48	304	2.245513
24	<i>pntA_</i>	0.35	0.536	0.525	510	2.222716
25	<i>sufC_</i>	0.287	0.564	0.507	248	2.245513
26	<i>ruvB_</i>	0.354	0.502	0.529	336	2.20412
27	<i>mglA_</i>	0.28	0.479	0.455	506	1.990783
28	<i>mgo_</i>	0.353	0.602	0.541	548	2.056905
29	<i>yfaA_</i>	0.296	0.544	0.512	562	2.068186
30	<i>menD_</i>	0.321	0.605	0.582	556	2.093422
31	<i>yfeR_</i>	0.269	0.525	0.54	308	2.245513
32	<i>rluD_</i>	0.393	0.505	0.529	326	2.245513
33	<i>ispD_</i>	0.274	0.528	0.556	236	2.222716
34	<i>epd_</i>	0.295	0.506	0.516	339	2.184691
35	<i>yggR_</i>	0.21	0.571	0.555	326	2.222716
36	<i>speC_</i>	0.313	0.555	0.523	711	2.056905
37	<i>glcB_</i>	0.394	0.581	0.535	723	2.056905
38	<i>ygiS_</i>	0.303	0.549	0.523	535	2.222716
39	<i>glnE_</i>	0.321	0.599	0.56	946	1.944483
40	<i>tdcB_</i>	0.379	0.425	0.465	329	2.184691
41	<i>nlpI_</i>	0.255	0.505	0.483	294	2.245513

42	<i>folP_</i>	0.332	0.524	0.518	282	2.245513
43	<i>arcB_</i>	0.356	0.555	0.506	778	1.999565
44	<i>yhdP_</i>	0.327	0.545	0.534	1266	1.810904
45	<i>gspA_</i>	0.206	0.445	0.49	489	2.068186
46	<i>malQ_</i>	0.329	0.591	0.552	694	2.173186
47	<i>glgA_</i>	0.31	0.555	0.55	477	2.20412
48	<i>glgB_</i>	0.387	0.566	0.533	728	1.999565
49	<i>yhiI_</i>	0.318	0.573	0.583	355	2.184691
50	<i>yhiN_</i>	0.32	0.616	0.552	400	2.167317
51	<i>bisC_</i>	0.329	0.535	0.545	777	1.999565
52	<i>dnaA_</i>	0.348	0.607	0.542	467	2.049218
53	<i>uvrD_</i>	0.387	0.678	0.578	720	1.966611
54	<i>yigL_</i>	0.391	0.573	0.518	266	2.20412
55	<i>rmuC_</i>	0.34	0.582	0.531	475	1.990783
56	<i>metL_</i>	0.396	0.645	0.584	810	1.937518
57	<i>oxyR_</i>	0.388	0.591	0.556	305	2.20412
58	<i>dusA_</i>	0.34	0.586	0.544	330	2.136721
59	<i>yjcE_</i>	0.323	0.645	0.559	549	2.20412
60	<i>hflX_</i>	0.336	0.549	0.54	426	2.222716
61	<i>ytfN_</i>	0.341	0.625	0.55	1259	2.017033
62	<i>htrE_</i>	0.27	0.376	0.434	865	1.982723
Lagging strand						
63	<i>ybfF_</i>	0.382	0.479	0.51	254	2.245513
64	<i>ltaE_</i>	0.328	0.594	0.564	333	2.167317
65	<i>poxB_</i>	0.341	0.574	0.541	572	2.093422
66	<i>ycgV_</i>	0.265	0.5	0.498	955	2.093422
67	<i>sapA_</i>	0.298	0.591	0.543	547	2.093422
68	<i>ydbK_</i>	0.372	0.53	0.533	1174	2.075547
69	<i>maoC_</i>	0.311	0.501	0.543	681	2.1959
70	<i>cfa_</i>	0.32	0.475	0.481	382	2.222716

71	<i>ynjE_</i>	0.373	0.517	0.536	435	2.136721
72	<i>yebT_</i>	0.327	0.562	0.537	877	1.974512
73	<i>amn_</i>	0.282	0.447	0.503	484	2.167317
74	<i>baeR_</i>	0.387	0.603	0.542	240	2.222716
75	<i>yehM_</i>	0.264	0.529	0.557	759	2.173186
76	<i>dld_</i>	0.365	0.583	0.523	571	1.974512
77	<i>yefF_</i>	0.233	0.513	0.514	529	2.025306
78	<i>dsdA_</i>	0.329	0.459	0.51	442	2.167317
79	<i>evgS_</i>	0.226	0.35	0.409	1197	1.91169
80	<i>pssA_</i>	0.346	0.534	0.49	451	2.245513
81	<i>yfjI_</i>	0.227	0.31	0.38	469	2.082785
82	<i>gabT_</i>	0.387	0.703	0.594	426	2.120574
83	<i>mutS_</i>	0.391	0.607	0.562	853	1.944483
84	<i>ygdH_</i>	0.362	0.57	0.515	454	2.209515
85	<i>acrF_</i>	0.309	0.56	0.516	1034	2.056905
86	<i>rsmB_</i>	0.295	0.427	0.497	429	2.149219
87	<i>yheS_</i>	0.399	0.614	0.554	637	2.0086
88	<i>mrcA_</i>	0.383	0.611	0.556	850	1.905256
89	<i>rfaF_</i>	0.326	0.538	0.537	348	2.184691
90	<i>waaA_</i>	0.324	0.529	0.54	425	2.184691
91	<i>spoT_</i>	0.383	0.601	0.538	702	1.974512
92	<i>cytR_</i>	0.317	0.543	0.522	341	2.222716
93	<i>trmA_</i>	0.373	0.58	0.51	366	2.167317
94	<i>iclR_</i>	0.288	0.534	0.543	274	2.222716
95	<i>lysC_</i>	0.383	0.544	0.54	449	2.120574
96	<i>plsB_</i>	0.377	0.636	0.552	807	2.158362
97	<i>alsA_</i>	0.289	0.523	0.48	510	2.056905
98	<i>ampC_</i>	0.262	0.485	0.503	377	2.167317
99	<i>hsdR_</i>	0.386	0.629	0.537	1188	2.056905
100	<i>lplA_</i>	0.335	0.677	0.579	338	2.184691

A Statistical-study on the nucleotide composition of bacterial chromosomes.

*Number of codons

Table 3.4: Strand specific mutational bias in codon 3rd position of HEG and WEG and in IR

	HEG		WEG		IR	
	LeS ₃	LaS ₃	LeS ₃	LaS ₃	LeS	LaS
A	0.108	0.112	0.125	0.141	0.290	0.308
T	0.392	0.383	0.219	0.228	0.300	0.295
G	0.272	0.279	0.385	0.331	0.216	0.191
C	0.228	0.227	0.270	0.300	0.194	0.206
Δ GC	-0.0073		0.0632		0.0457	
Δ AT	-0.0102		-0.0187		-0.0192	

HEG: highly expressed genes; frequency of the nucleotides at the 3rd position of genes studied.

WEG: weakly expressed genes; frequency of the nucleotides at the 3rd position of genes studied.

IR: intergenic regions; abundance of the nucleotides in the IR. The IR > 100 nucleotides size were only considered. The abundance values of the nucleotides were calculated excluding 50 nucleotides from each ends (5' as well as 3') of an IR.

LeS₃: 3rd position of codons in the leading strand; LaS₃: 3rd position of codons in the lagging strand

Δ GC: is the average difference of GC skews $[(G-C)/(G+C)]$ between the strands (LeS - LaS);

Δ AT: is the average difference of AT skews $[(A-T)/(A+T)]$ between the strands (LeS - LaS)

3.4.3. ATS₃ and GCS₃ between LeS and LaS

The 3rd position of a family box codon has the maximum degeneracy. So we limited our analysis to the codons of the eight family boxes in the genetic code for studying the effect of strand-specific mutational bias on different codons. AT-skew (ATS₃) and GC-skew (GCS₃) at the 3rd position of codons in different family boxes were found out. The difference of the skews between the strands was calculated in both types of genes (Table 3.5). The significance of GCS₃ difference was observed in seven family boxes, except CTN_WEG, in

case of weakly expressed genes whereas the same was observed to be significant in only one family box, GCN_HEG (Table 3.5). This observation is in further support of the view that strand-specific mutational bias influences codon usage bias of weakly expressed genes to a greater extent than that of highly expressed genes.

3.4.4. Higher CRSCU in case of weakly expressed genes than highly expressed genes

We estimated the change in relative synonymous codon usage (CRSCU) in each family box. CRSCU was found to be higher in weakly expressed genes than in highly expressed genes (Table 3.6). In case of highly expressed genes, the maximum CRSCU was found in CCN_pro (0.118) family box and the minimum was found in CTN_leu (0.014) family box. In case of weakly expressed genes, the maximum CRSCU was found in GCN_ala (0.176) family box and the minimum was found in CTN_leu (0.054) family box. CRSCU of weakly expressed genes was found to be significantly higher than highly expressed genes (p value < 0.001 , one tail t-test). The correlation coefficient between CRSCU for the two types of genes (highly and weakly) was significant ($r = 0.80$). This indicates that selection is effective in both types of genes though it is more in highly expressed genes. A weak negative correlation was observed between the CRSCU and tRNA ratio (Satapathy *et al.*, 2010) for highly expressed genes ($r = -0.65$) and weakly expressed genes ($r = -0.4249$). This indicates that stronger the selection on a family box lesser the CRSCU.

Table 3.5: Strand specific $ATS_3 = [A_3/(A_3+T_3)]$ and $GCS_3 = [G_3/(G_3+C_3)]$ in highly expressed genes (HEG) and weakly expressed genes (WEG)

	<i>LeS_HEG</i>	<i>LaS_HEG</i>		<i>LeS_WEG</i>	<i>LaS_WEG</i>	
Family box	$A_3/(A_3+T_3)$	$A_3/(A_3+T_3)$	p value*	$A_3/(A_3+T_3)$	$A_3/(A_3+T_3)$	p value*
ACN	0.110	0.100	0.418	0.408	0.540	0.000
CCN	0.489	0.698	0.002	0.540	0.526	0.343
CGN	0.011	0.017	0.440	0.138	0.126	0.320
CTN	0.114	0.139	0.397	0.277	0.277	0.496
GCN	0.418	0.401	0.263	0.606	0.561	0.026
GGN	0.025	0.029	0.459	0.214	0.248	0.085
GTN	0.328	0.296	0.151	0.363	0.364	0.491
TCN	0.095	0.128	0.242	0.434	0.480	0.098
	$G_3/(G_3+C_3)$	$G_3/(G_3+C_3)$	p value	$G_3/(G_3+C_3)$	$G_3/(G_3+C_3)$	p value
ACN	0.147	0.122	0.269	0.441	0.396	0.022
CCN	0.983	0.978	0.452	0.854	0.768	0.000
CGN	0.022	0.011	0.426	0.211	0.157	0.013
CTN	0.926	0.933	0.396	0.815	0.804	0.277
GCN	0.690	0.608	0.012	0.623	0.528	0.000
GGN	0.052	0.067	0.345	0.303	0.237	0.001
GTN	0.673	0.688	0.377	0.677	0.587	0.000
TCN	0.126	0.081	0.217	0.570	0.507	0.026

A Statistical study on the nucleotide composition of bacterial chromosomes.

*Z test for testing significance of the difference of proportions $A_3/(A_3+T_3)$ and $G_3/(G_3+C_3)$ in the LeS and LaS. Significant p-values ($p < 0.05$; Z-test) are shown in bold.

Table 3.6: Change in relative synonymous codon usage (CRSCU) between the strands

Family box	HEG	WEG
GCN_ala	0.091	0.176
CGN_arg	0.075	0.078
GGN_gly	0.035	0.087
CTN_leu	0.014	0.054
CCN_pro	0.118	0.162
TCN_ser	0.054	0.116
ACN_thr	0.059	0.129
GTN_val	0.060	0.130

CRSCU in weakly expressed genes (WEG) are significantly more ($p < 0.001$ one tail t-test) than highly expressed genes (HEG)

3.4.5. Higher SCF in case of weakly expressed genes than highly expressed genes

Synonymous codon frequencies of family box codons were compared between the strands (Table 3.7). In case of highly expressed genes, out of the 32 codons analyzed, synonymous codon frequencies of 11 codons were found to be significantly different between the strands. In case of weakly expressed genes, out of 32 codons analyzed, 19 codons were found to be significantly different between the strands. We compared the synonymous codon frequencies between the strands in the context of strand-specific mutational bias i.e. synonymous codon frequencies of G/T ending codons will be higher in LeS than LaS and the reverse for the codons ending with A/C. Out of the above eleven codons observed in case of highly expressed genes, synonymous codon frequencies of five codons were not in accordance with the strand-specific mutational bias assumption. These are ACT, GCT, GGC,

GTA and GTT. Out of the above nineteen codons observed in case of weakly expressed genes, synonymous codon frequencies of only three codons were not in accordance with the strand-specific mutational bias assumption. These are CCT, CTT and GCT. This indicates that strand-specific mutational bias is an influential factor for the observed difference in synonymous codon frequencies between the strands for weakly expressed genes.

Table 3.7: Synonymous codon frequency of family box codons in highly expressed genes and weakly expressed genes

Sl.	Codon	HEG			WEG		
		LeS	LaS	p-value*	LeS	LaS	p-value*
1	ACA_thr	0.043	0.042	0.400	0.103	0.156	0.000
2	ACC	0.523	0.513	0.209	0.418	0.430	0.141
3	ACG	0.090	0.071	0.064	0.330	0.282	0.000
4	ACT	0.344	0.374	0.008	0.149	0.133	0.061
5	CCA_pro	0.125	0.180	0.000	0.180	0.192	0.148
6	CCC	0.013	0.016	0.399	0.097	0.148	0.000
7	CCG	0.731	0.725	0.351	0.569	0.488	0.000
8	CCT	0.131	0.078	0.000	0.154	0.172	0.043
9	CGA_arg	0.008	0.011	0.377	0.065	0.057	0.225
10	CGC	0.309	0.342	0.003	0.419	0.458	0.000
11	CGG	0.007	0.004	0.398	0.112	0.085	0.003
12	CGT	0.677	0.643	0.002	0.405	0.400	0.306
13	CTA_leu	0.008	0.010	0.408	0.053	0.060	0.217
14	CTC	0.069	0.062	0.255	0.150	0.154	0.334
15	CTG	0.861	0.863	0.441	0.659	0.632	0.001
16	CTT	0.061	0.064	0.391	0.138	0.155	0.022
17	GCA_ala	0.257	0.253	0.310	0.202	0.213	0.083
18	GCC	0.119	0.145	0.002	0.251	0.293	0.000
19	GCG	0.266	0.224	0.000	0.415	0.327	0.000

20	GCT	0.358	0.378	0.010	0.131	0.167	0.000
21	GGA_gly	0.014	0.016	0.407	0.090	0.110	0.015
22	GGC	0.418	0.400	0.036	0.404	0.425	0.009
23	GGG	0.023	0.029	0.268	0.176	0.132	0.000
24	GGT	0.545	0.555	0.172	0.331	0.333	0.388
25	GTA_val	0.223	0.204	0.030	0.142	0.149	0.226
26	GTC	0.105	0.097	0.200	0.197	0.244	0.000
27	GTG	0.216	0.213	0.372	0.413	0.347	0.000
28	GTT	0.456	0.486	0.001	0.249	0.260	0.107
29	TCA_ser	0.056	0.077	0.084	0.183	0.223	0.002
30	TCC	0.361	0.367	0.347	0.249	0.264	0.129
31	TCG	0.052	0.032	0.098	0.330	0.272	0.000
32	TCT	0.532	0.524	0.315	0.239	0.241	0.434

* Significant *p*-values ($p < 0.05$) are shown in bold.

3.5. Discussion

Our hypothesis in this work is that the influence of strand-specific mutational bias on CUB varies among genes within a genome. In concordance to this, strand-specific mutational bias is found influencing CUB in weakly expressed genes to a greater extent than that in highly expressed genes in *E. coli*. This observation is important because a specific mutational bias, which is associated with replication, has been demonstrated to be increased in case of weakly expressed genes in an organism with strong selected codon usage bias. The observation in *E. coli* supports the view that under the same mutational pressure, selection on codon usage varies depending upon the expression levels of genes in a genome. The differential effect of strand-specific mutational bias on CUB in highly expressed genes and weakly expressed genes can mainly be attributed to the effect of purifying selection in *E. coli*. The negative correlation between CRSCU and tRNA ratio is in support of this. In addition, the high CRSCU as well as GCS₃ in case of weakly expressed genes are in favour of the explanation of purifying selection.

Two other arguments favoring higher mutation in weakly expressed genes than highly expressed genes due to transcription might be given to explain the observations. First, the activity of transcription coupled repair on template strand is directly proportional to number of times the transcript is made on the strand. This makes a higher probability of fixing a mutation in weakly expressed genes than in highly expressed genes. However, transcription coupled repair is mainly effective against DNA lesions such as pyrimidine dimers that are known to cause C→T transition on the template DNA (Francino and Ochman, 2001). This suggests a higher abundance of A at the 3rd position of codons in weakly expressed genes than in highly expressed genes. In contrast to this we observed a strand specific codon usage bias in case of weakly expressed genes, which goes against the first explanation. Second, the non-template DNA undergoes higher deamination of cytosine nucleotides owing to its temporary single stranded condition (Francino and Ochman, 1997). This suggests that a higher increase in T at the 3rd position of codons in highly expressed genes than in weakly expressed genes due to C→T transition in the non-template DNA. This is also going to happen in LeS as well as LaS. Therefore, the asymmetric deamination between the strands can not be the correct explanation for the strand specific codon usage in case of weakly expressed genes observed in this study.

The compositional asymmetry between the strands is known to affect the amino acid composition between the strands in different bacteria (Rocha *et al.*, 1999; Mackiewicz *et al.*, 1999). The best example is found in *Borrelia burgdorferi*, between the homologous genes (BB0629 and BB0408) that are highly similar to the *E. coli fruA*, one (BB0408) is present on the LeS and the other (BB0629) is present on the LaS (Rocha *et al.*, 1999). The two proteins reveal a very strong mutational polarization when non-synonymous substitution compared between valine and isoleucine codons in both the strands (Rocha *et al.*, 1999). In *E. coli*, the proteome compositions between LeS and LaS are not different with respect to highly expressed genes as well as weakly expressed genes (data not shown). This suggests that the influence of strand-specific mutational bias on weakly expressed genes in *E. coli* is limited only to synonymous changes.

Daubin and Perrière had earlier shown in *E. coli* and other prokaryotes that there is G+C₃ structuring along the chromosome: a tendency toward an A+T enrichment near

replication terminus (Daubin and Perrière, 2003). The A+T enrichment towards the terminus has arisen mainly due to mutation pressure (Ochman, 2003), which leads to the observation of positional constraints on gene orientation, length and codon usage in bacterial chromosomes with regard to the position of replication origin and terminus (Arakawa and Tomita, 2007). The effect G+C₃ structuring on codon usage separately for highly expressed genes and weakly expressed genes have not been analyzed in the above studies. We did an analysis to compare the effect of the A+T enrichment on the codon usage in the two gene types. Similar to the strand-specific mutational bias result, A+T₃ enrichment was found to be significantly ($p \ll 0.0001$) effective on weakly expressed genes towards the terminus in comparison to the origin whereas its effect on highly expressed genes is insignificant. G+C₃ structuring in genomes is a direct effect of replication because the bias is observed from replication origin to terminus. Therefore, its differential effect on weakly expressed genes in *E. coli* is attributed to replication and not to any transcription effect. The observation here is in support of the view that mutation influences weakly expressed genes to a greater extent than highly expressed genes due to higher purifying selection in the latter.

It is pertinent to note that several studies have been done earlier to study strand-specific mutational bias between the strands with respect to replichores, 3rd position of codons and intergenic regions (Lobry, 1996; Mclean *et al.*, 1998; Lobry and Sueoka, 2002). So far our knowledge is concerned, this is the first report with respect to the differential influence of strand-specific mutational bias on highly and weakly expressed genes in *E. coli* chromosome. However, the result we find here in *E. coli* might be different in other bacteria because (i). the strength of selected codon usage bias have been reported to be different among bacteria, (ii). the strand specific mutational bias varies among bacteria (Mclean *et al.*, 1998; Lobry and Sueoka, 2002; Morton and Morton, 2007) and (iii) selection effects on the positioning of genes and gene structures in bacterial genomes is different in regions surrounding the terminus of replication from the rest of the genome and these positional effects are partly attributed to the A+T enrichment near the terminus (Arakawa and Tomita, 2007). So, future research with the availability of more proteome data in other prokaryotic genomes (e.g. *Bacillus subtilis* etc) will give more insight into the influence of strand specific mutational bias on highly and weakly expressed genes in these organisms.

CHAPTER IV

4. Selected codon usage bias in bacterial chromosomes

4.1. Abstract

Codon usage bias exhibited in the coding sequences reflects combined effects of mutation and selection where both of them are confounded. A new index for unevenness of codon usage in genes has been developed in this study. The index is named as unevenness of codon usage (UCU). UCU in a gene 'g' [UCU(g)] is the amount of average variation of synonymous codon frequency in eight family boxes measured for each third position letter (A, C, G, T). UCU(g) exhibited significant positive correlation with gene expression in *Escherichia coli* and *Saccharomyces cerevisiae*. This indicated that UCU(g) measures selected codon usage bias in these organisms. UCU(g) was studied in 76 bacterial genomes and compared with principal axis of correspondence analysis as well as with effective number of codons. 20 out of 76 genomes exhibited strong correlation, 19 out of 76 exhibited weak correlation with the principal axis of correspondence analysis. Most of these 39 genomes also exhibited significant correlation between UCU(g) and effective number of codons (ENc). These results supported the assumption that the new index measures the selected codon usage bias in genes. UCU(g) might be of interest to molecular evolutionary biologists as it is not dependent upon the gene expression data for finding the selected codon usage bias in genes.

4.2. Introduction

Though synonymous codons encode the same amino acid, they are used with different frequencies. This is known as codon usage bias, which is a consequence of both mutation and selection pressures (here onwards called mutation and selection) in unicellular and multicellular organisms (Sharp *et al.*, 1995; Ermolaeva, 2001; Hershberg and Petrov, 2008; Yang and Nielsen, 2008). In an organism, major factors known to influence codon usage bias are (i). genomic G+C content (Muto and Osawa, 1987; Chen *et al.*, 2004), (ii). strand-specific mutational bias (Lobry, 1996; McInerney, 1998; Frank and Lobry, 1999; Powdel *et al.*, 2010) and (iii). gene expression (Ikemura, 1985; Duret and Mouchiroud, 1999; Hiraoka *et al.*, 2009). The first and the second factors belong to mutation whereas the third factor belongs to selection. The extent or magnitude of these factors varies greatly among species

(Sharp *et al.*, 2005). The major challenge for molecular evolutionary biologists is to estimate the selection responsible for codon usage bias in a gene (dos Reis and Wernisch, 2009).

The two initial observations suggesting the role of selection on codon usage bias in a gene are as follows. First, the abundance values of different tRNA molecules are not same in the cytosol and there is positive correlation between the tRNA abundance values with the codon usage in organisms (Ikemura, 1981). Second, the selection on codon usage bias is unidirectional in organisms i.e. expression is positively correlated with the selection for optimal codons in genes for efficient translation (Sharp and Li, 1986a).

Based on the above two points, a method has been developed in this study to measure unevenness of codon usage in genes. The assumption in the method is as follows. For different amino acid codons within a gene, the magnitude and direction of mutation on codon usage bias remain same while the magnitude and direction of selection vary from one amino acid codons to other amino acid codons. The differential selection is due to tRNA molecules, whose abundance values differ inside a cell for different amino acid codons (Ikemura, 1981, 1985). In a hypothetical condition with zero selection, and codon usage bias being determined only by mutation, the pattern of codon usage bias for different amino acids will be similar: If frequencies of the four glycine codons (GGN) in a gene is such that $F_{GGA} > F_{GGT} > F_{GGC} > F_{GGG}$ (F_{XYZ} : frequency of the codon XYZ), then the frequencies among the four alanine codons (GCN) in the gene is likely to be $F_{GCA} > F_{GCT} > F_{GCC} > F_{GCG}$. The same pattern will also be observed in other family boxes in the gene i.e. frequencies of 'A' ending codons are the highest and the same of 'G' ending codons are the lowest among the four synonymous codons, in different family boxes. The invariable pattern of codon usage among different family boxes will also be observed if selection remains same for different amino acid codons. However, this is not observed in a cell because the chance of either zero or invariable selection on codons in an organism seems very less. Codon usage bias patterns vary among different amino acids because the selection on different amino acid codons is variable and independent of one another: if selection on four glycine codons causes their frequencies in a gene to be like $F_{GGC} > F_{GGT} > F_{GGA} > F_{GGG}$, selection on four alanine codons in the gene might cause their frequencies to be like $F_{GCA} > F_{GCT} > F_{GCC} > F_{GCG}$. In other family box codons the patterns may also be different. In short, variation of frequencies among codons

ending with the same nucleotide in the eight family boxes (F_{ACA} , F_{CGA} , F_{GCA} , F_{GGA} etc) is going to be less in genes with weak selection and the same is going to be high in genes with strong selection. We have used this logic to measure the variation of codon usage bias pattern among different family box codons in genes. The mean value of the “average deviations measured for the synonymous codon frequencies ending with same nucleotide” is named as “Unevenness of codon usage (UCU)”. $UCU(g)$ correlated significantly with expression in *Escherichia coli* (Ishihama *et al.*, 2008) and *S. cerevisiae* (Ghaemmaghami *et al.*, 2003). Comparing with correspondence analysis and effective number of codons it was found out that $UCU(g)$ indeed represents the selected codon usage bias in genes. $UCU(g)$ might be of interest to molecular evolutionary biologist as it does not require the previous knowledge of gene expression to measure the selected codon usage bias in a gene.

4.3. Materials and Methods

4.3.1. Calculations for $UCU(g)$

The formula for uneven codon usage of a gene ‘g’, $UCU(g)$ is defined as follows:

$$UCU(g) = \left[\frac{1}{4} \sum_{z \in \{A,T,G,C\}} \left\{ \frac{1}{n} \sum_{yz \in \{AC,GG,GC,GT,CC,CG,CT,TC\}} |SCF_{yz}^F - M_z| \right\} \right] / k$$

SCF_{yz}^F is synonymous codon frequency (SCF) of a codon ‘xyz’ within a family box

(F). For example: $SCF_{ACA}^F = \frac{X_{ACA}}{\sum_{N \in \{A,T,G,C\}} X_{ACN}}$

Where X_{ACA} is the number of occurrences of ACA codon and $\sum_{N \in \{A,C,G,T\}} X_{ACN}$ is the total number of occurrences of ACN family box codons within a gene sequence. For the eight family box codons in the genetic code { ACN (thr), CCN (pro), CGN (arg), CTN (leu), GCN (ala), GGN (gly), GTN (val), TCN (ser)} there will be 32 such SCF values. SCF is a modification of the commonly used relative synonymous codon usage (RSCU) value (Sharp *et al.*, 1986).

M_z in the above equation represents the average (mean) of SCF_{xyz}^f values among all family box codons with 'z' at the third position {z= (A, T, G, C)}.

'UCU(g)' is the mean of the four average deviations. The average deviations were calculated for a set of SCF values ending with same nucleotide. The divisor n is representing the number of family boxes used to calculate UCU(g) and it is a measure depending on a maximum of 32 deviations grouped into four. This will result a maximum value of n equal to 8. 'k' is a constant depending on number of family boxes used in UCU(g). 'k' is hypothetical value of uneven codon usage measure in case of highly biased gene using only the optimal codon in each family box and it is variable with respect to the number of family boxes used. The average deviation value is scaled to 'k' to make it comparable in the [0,1] range.

CAI values were calculated using the program CodonW (<ftp://molbiol.ox.ac.uk/cu/codonW.tar.Z>) (Peden, 1999). Gene sequences were downloaded from DDBJ (www.gib.genes.nig.ac.jp) sites. Gene expression data for *E. coli* and *S. cerevisiae* were taken from Ishihama *et al.*, (2008), and Ghaemmaghami *et al.*, (2003), respectively. CodonW was also used for correspondence analysis of genes in bacterial genomes to find the major source of codon variation. The effective number of codons (ENc) was also calculated using CodonW. The number of expected effective number of codons $f(\theta_g)$ was found out using the formula given in the section 1.4.3.3.

4.3.2. Calculation of strand specific mutational bias in the intergenic regions (mut_ir) as well as at the 3rd position of codons (mut_3)

Nucleotide sequences of the intergenic regions (IR) in bacterial chromosomes were downloaded from the comprehensive microbial resources (CMR) web site (<http://cmr.jcvi.org/>). The coordinate points of replication origin and terminus were taken from the URL: <http://pbil.univ-lyon1.fr/datasets/Necsulea2007/html/index.html>, (Necsulea and Lobry, 2007). IR of size at least 200 nucleotides in a chromosomes were considered for calculating the strand specific mutational bias. Fifty nucleotides from the 5' end and equal number of nucleotides from the 3' end were removed before analysis. The removal of the 100 nucleotides was done to consider regions that are relatively more away from the gene and

therefore more neutral. These IR then were divided into two groups: intergenic regions of leading strands (LeS_ir) and intergenic regions of lagging strands (LaS_ir). GCS = $G/(G+C)$ and ATS = $A/(A+T)$ were calculated separately for the LeS_ir and LaS_ir based on total compositional abundance values of the nucleotides in each group. So strand specific mutational bias in the intergenic region (mut_ir) is as follows.

$$\text{mut_ir} = |\text{GCS}_{\text{LeS}} - \text{GCS}_{\text{LaS}}|$$

Using similar procedure, strand specific mutational bias at the 3rd position of codons (mut_c3) was calculated as follows:

$$\text{mut_c3} = |\text{GCS3}_{\text{LeS}} - \text{GCS3}_{\text{LaS}}|$$

$$\text{GCS3} = G3/(G3+C3) \text{ and } \text{ATS3} = A3/(A3+T3)$$

X3 defined as the abundance values of X at the 3rd position of codons {X= (A, C, G, T)}.

4.4. Results

4.4.1. Uneven codon usage measure of genes [UCU(g)] in *E. coli* as well as in *S. cerevisiae*

UCU(g) of 3727 *E. coli* genes (having ≥ 100 codons) ranges from 0.130 (*yeeO*) to 0.664 (*erpA*). The average UCU(g) value is 0.336 with standard deviation 0.075. Several genes known to be highly expressed were found with high UCU(g). Few examples are *rplL* [UCU(g) 0.653 is the 2nd highest], *rplK* [UCU(g) 0.597 is the 7th highest], *rpsA* [UCU(g) 0.577 is the 14th highest], and chaperonin *groL* [UCU(g) 0.572 is the 17th highest]. UCU(g) values are high for *E. coli* genes encoding ribosomal proteins (*rps*, *rpl*), outer membrane proteins (*omp*), elongation factors (*tuf*) and RNA polymerase subunits (*rpo*). The general observation of high UCU(g) for highly expressed genes is in concordance with the idea of high selection on these genes.

UCU(g) of 6238 *S. cerevisiae* genes ranges from 0.062 (YKL202W) to 0.665 (YGL102C) (out of 7080 coding sequences, 841 genes with < 100 codons and 01 gene with insufficient family box codons were not studied). The average UCU(g) is 0.256 with standard deviation 0.084. Uneven codon usage of several genes encoding large (*rpl*) and small (*rps*) ribosomal proteins, translation elongation factors (*efb1*, *eft1*, *tef1*, *tef2*) are among the top 120 UCU(g) values. Two genes coding for ribosomal proteins, *rpl35B* and *rpl35A*, have 6th and

7th highest UCU(*g*) respectively. The *pau* genes that constitute the largest multigene family in yeast and are known to be induced in anaerobiosis (Rachidi *et al.*, 2000), are observed among the genes with high UCU(*g*).

Codon adaptation index (CAI; Sharp and Li, 1987b) is widely used to predict gene expression (Henry and Sharp, 2007), which correlates positively with codon usage bias in *E. coli* and *S. cerevisiae*. Significant correlation between CAI and gene expression in *E. coli* (dos Reis *et al.*, 2003) and *S. cerevisiae* (Coghlan and Wolfe, 2000) has already been reported. So we did a correlation study between UCU(*g*) and CAI. Significant correlation was observed between UCU(*g*) and CAI values in *E. coli* (Pearson $r = 0.711$) as well as in *S. cerevisiae* (Pearson $r = 0.443$) (Table 4.1; Fig. 4.1). The correlation was stronger for larger genes than smaller genes. In *E. coli*, the correlation coefficient was 0.925 in case of genes with ≥ 600 codons (Table 4.1) and the same was 0.940 in case of genes with ≥ 800 codons. Similarly, in *S. cerevisiae*, the correlation coefficient (Pearson r) was 0.751 in case of genes with ≥ 800 codons.

The expression levels of many genes are known from the proteome analysis results in *E. coli* (Ishihama *et al.*, 2008) and *S. cerevisiae* (Ghaemmaghami *et al.*, 2003). We did a correlation between UCU(*g*) and gene expression, and compared it with the correlation between CAI and gene expression in these two organisms. With the expression values of 894 genes (\log_2 protein abundance; Ishihama *et al.*, 2008) that encode cytosolic proteins in *E. coli*, UCU(*g*) exhibited strong positive correlation (Table 4.2, Fig. 4.2). The correlation was stronger in case of larger genes (Table 4.2). The correlation coefficient was 0.717 in case of genes with ≥ 800 codons (Table 4.2). With the expression of 3758 yeast genes (\log_2 protein abundance; Ghaemmaghami *et al.*, 2003), UCU(*g*) exhibited significant positive correlation (Table 4.2; Fig. 4.2). Like *E. coli*, in *S. cerevisiae* the correlation was stronger in case of larger genes. The Pearson correlation coefficient (r) was 0.589 in case of genes with ≥ 800 codons. In both the organisms, the correlation is stronger in case of larger genes than smaller genes because the number of family box codons available is more in the former than the latter. This was proved by studying only the genes where UCU(*g*) values have been calculated using all the eight family boxes. As expected, the correlation coefficient was better (data not shown). Further we arranged the genes according to the abundance values of family

box codons instead of their size and then did a correlation analysis. The correlation was better when genes were arranged according to the abundance values of family box codons (data not shown). The positive correlation of $UCU(g)$ with CAI as well as with gene expression indicated that $UCU(g)$ measures the selected codon usage bias in genes of these two organisms.

Table 4.1: Correlation between $UCU(g)$ and CAI in *E. coli* and *S. cerevisiae* (yeast)

Sl no	Size ^a	<i>E. coli</i>		Yeast	
		No. of genes	^b r(UCU,CAI)	No. of genes	^b r(UCU,CAI)
1	≥100	3727	0.711	6238	0.443
2	≥200	2843	0.798	4806	0.595
3	≥300	1894	0.853	3886	0.671
4	≥400	1113	0.885	3000	0.706
5	≥500	569	0.913	2241	0.727
6	≥600	338	0.925	1629	0.725
7	≥700	223	0.934	1195	0.753
8	≥800	134	0.940	905	0.751

^aGene size in terms of total number of codons.

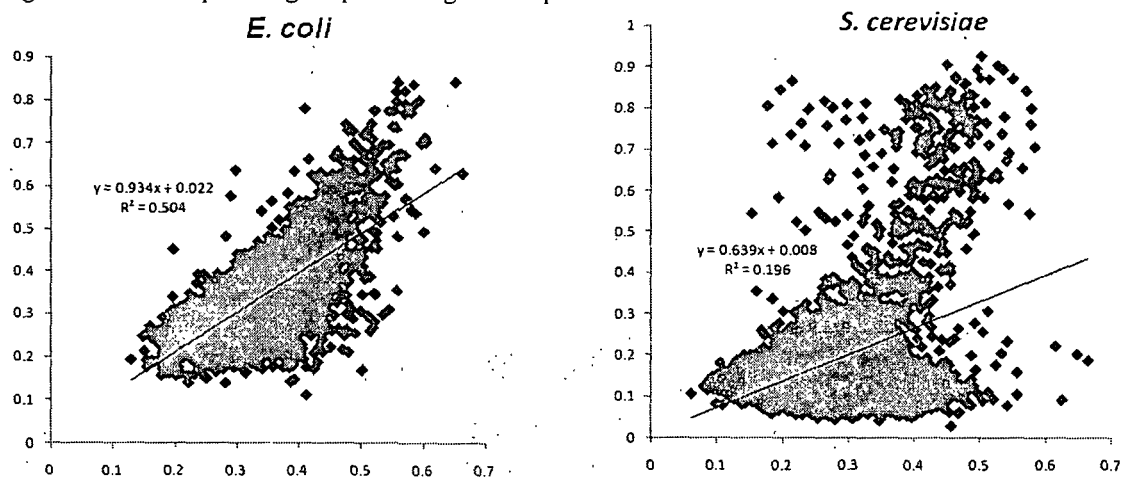
^bPearson correlation coefficient (r) between $UCU(g)$ and CAI. All r -values are statistically significant ($p < 0.0001$).

Table 4.2: Correlation of UCU(g) and CAI with gene expression in *E. coli* and *S. cerevisiae* (yeast)

Size ^a	<i>E. coli</i>			yeast		
	No. of genes	^b r(UCU,exp)	^c r(CAI,exp)	No. of genes	^b r(UCU,exp)	^c r(CAI,exp)
≥100	894	0.624	0.712	3758	0.434	0.573
≥200	708	0.651	0.729	3248	0.482	0.548
≥300	491	0.662	0.736	2664	0.523	0.550
≥400	312	0.704	0.778	2069	0.536	0.559
≥500	187	0.728	0.799	1550	0.571	0.588
≥600	127	0.749	0.816	1150	0.574	0.603
≥700	90	0.767	0.824	848	0.573	0.608
≥800	59	0.717	0.767	625	0.589	0.599

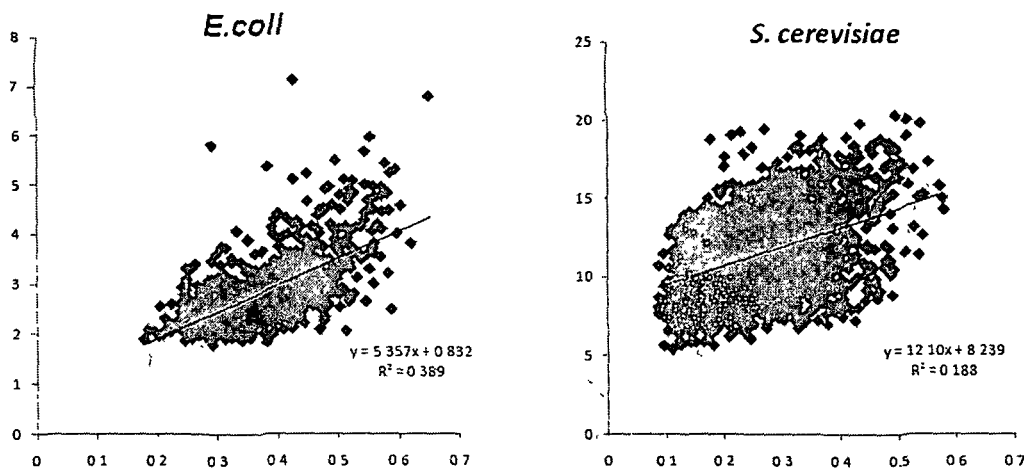
^aGene size in terms of total number of codons. ^bPearson correlation coefficient (r) between UCU(g) and CAI. All r -values are statistically significant ($p < 0.0001$).

Figure 4.1: A two panel figure presenting scatter plots of UCU(g) vs. CAI



Scatter plots of UCU(g) vs. CAI in *E. coli* (left) as well as in *S. cerevisiae* (right). X-axis and Y-axis define UCU(g) and CAI values, respectively. The Pearson correlation coefficients are significant in both the cases.

Figure 4.2: A two panel figure presenting scatter plot of UCU(g) vs. gene expression



Scatter plots of UCU(g) vs. \log_2 protein abundance in *E. coli* (left) as well as in *S. cerevisiae* (right). X-axis and Y-axis define UCU(g) and \log_2 protein abundance respectively. Gene expression data for *E. coli* and *S. cerevisiae* are taken from the protein abundance values reported in Ishihama *et al.*, (2008), and Ghaemmaghami *et al.*, (2003), respectively. The Pearson correlation coefficients are significant in both the cases.

4.4.2 UCU(g) correlates with primary axis of correspondence analysis as well as with effective number of codons (ENC) in several bacteria

Correspondence analysis is usually used to identify major sources of variation in synonymous codon usage among genes and provides a way to identify highly expressed genes (Suzuki *et al.*, 2008). We did correlation between UCU(g) (gene size ≥ 500 codons) and first three axes of correspondence analysis in different bacteria. The list of bacteria was taken from Sharp *et al.* (2005) where the strength of selected codon usage bias 'S' had been estimated by a population genetics model (Sharp *et al.*, 2005). In bacteria with high 'S' such as *Bacillus anthracis* Ames, *Corynebacterium glutamicum*, *E. coli*, *Lactococcus lactis lactis* etc high correlation was observed between UCU(g) and the primary axis (Table 4.3). In bacteria with low 'S' such as *Borrelia burgdorferi* B31, *Buchnera aphidicola* Sg, *Treponema pallidum*, *Xylella fastidiosa* 9a5c etc the correlation between UCU(g) and primary axis was very weak. The 76 bacteria in the list were grouped into three categories depending upon the Pearson r value between UCU(g) and the primary axis: category 1 includes 20 bacteria

(having the Pearson r) $|r| \geq 0.6$, category 2 includes 19 bacteria with $0.6 < |r| \geq 0.4$, and category 3 includes 37 bacteria with $|r| < 0.4$ (Table 4.3). Out of 20 bacteria in category 1, 7 (35%) bacterial genome GC% was below 50. Out of 19 bacteria in category 2, 11 (58%) bacterial genome GC% was below 50. Out of 37 bacteria in category 3, 30 (81%) bacterial genome GC% was below 50. This indicated that AT rich genomes (Table 4.3) having a higher probability of belonging to category 3.

We calculated the effective number of codons (ENc) using the principle of Wright (1990). The objective behind this exercise is to see whether UCU(g) is correlated with ENc. The expected $N_c [f(\theta_g)]$ under random mutation in 3rd codon position were also calculated using the relation defined in the section 1.4.3.3. of Chapter I. The correlation result is also presented in Table 4.3. In most of the bacteria in the list, UCU(g) and ENc are found to be negatively correlated. This is in concordant with the theory based on which UCU(g) is calculated. Bacteria that have been characterized under category 1 and 2 (Table 4.3), most of them exhibited high correlation (Pearson $|r|$) between N_c and UCU(g). Some of the examples are *Bifidobacterium longum*, *Corynebacterium glutamicum*, *E. coli*, *Pseudomonas putida*, *Pseudomonas syringae*, *Salmonella enterica*, *Shewanella oneidensis*, *Xanthomonas campestris*, etc. In bacteria such as *Streptomyces coelicolor*, *Xylella fastidiosa*, with genome GC% > than 50 did not exhibit high correlation. These bacteria are listed under the category 3. Usually the correlation between ENc and UCU(g) was low in genomes listed under category 3 and most of them are AT rich genomes.

The first bacterial genome, *B. burgdorferi*, known with very weak selected codon usage bias was also with high strand specific mutational bias. So we studied the strand specific mutational bias in genomes with variable selected codon usage bias in bacterial genomes. *Clostridium perfringens*, with strong selected codon usage bias (high 'S' value) (Sharp *et al.*, 2005) was found with high strand specific mutational bias. Usually, Firmicutes were found with high strand specific mutational bias. However, most of the bacteria in this group were reported to have high selected codon usage bias (Sharp *et al.*, 2005). *Corynebacterium glutamicum*, (Actinobacteria group) with high strand specific mutational bias was reported with high selected codon usage bias. This indicated that strand specific

mutational bias might not be the deciding factor for the selected codon usage bias in a genome. In the category 1, the strand specific mutational bias was found to cover a wide range from 0.025 to 0.133 (Table 4.3). Out of 20 bacteria, 6 bacterial genomes exhibited high strand specific mutational bias $mut_ir > 0.1$. So strand specific mutational bias is likely not the determining factor of selected codon usage bias.

Table 4.3: Correlation analysis of UCU(g) in 76 bacterial genomes

Sl	Name	GC %	r1	r2	r3	mut_3	mut_ir	Category	Phylotype	rUCU(g)ExpNc	rUCU(g)Nc
1	<i>Corynebacterium glutamicum</i> ATCC 13032 Bielefeld	54	0.722	0.494	-0.125	0.058	0.107	1	Actinobacteria	0.462	-0.793
2	<i>Corynebacterium efficiens</i> YS-314	63	0.666	-0.423	-0.269	0.004	0.084	1	Do	-0.187	-0.702
3	<i>Bifidobacterium longum</i> NCC2705	60	-0.742	0.399	0.124	0.013	0.023	1	Do	-0.323	-0.754
4	<i>Mycobacterium tuberculosis</i> H37Rv (lab strain)	66	0.445	-0.166	0.068	0.048	0.032	2	Do	-0.358	-0.53
5	<i>Mycobacterium leprae</i>	58	-0.45	-0.066	0.2	0.063	0.053	2	Do	0.256	-0.617
6	<i>Streptomyces coelicolor</i> A3(2)	72	0.196	0.054	-0.052	0.007	0.002	3	Do	-0.187	-0.286
7	<i>Tropheryma whipplei</i> strTwist	46	0.013	0.231	0.025	0.038	0.118	3	Do	-0.065	-0.374
8	<i>Streptomyces avermitilis</i> MA-4680	71	-0.185	-0.063	0.143	0.03	0.027	3	Do	-0.129	-0.29
9	<i>Aquifex aeolicus</i> VF5	43	-0.46	-0.266	-0.26	0.004	0.019	2	Aquificae	0.274	-0.639
10	<i>Bacteroides thetaiotaomicron</i> VPI-5482	42	-0.47	0.323	0.484	0.131	0.081	2	Bacteroidetes	0.388	-0.433
11	<i>Chlorobium tepidum</i> TLS	57	-0.326	-0.061	0.468	0.05	0.08	3	Do	-0.034	-0.512
12	<i>Chlamydia trachomatis</i>	41	0.11	0.071	-0.059	0.222	0.15	3	Chlamydiae	0.056	-0.371
13	<i>Chlamydia muridarum</i> strain Nigg	40	0.011	0.064	0.144	0.245	0.15	3	Do	0.083	-0.384

14	<i>Chlamydophila caviae</i> GPIC	39	-0.013	0.053	0.188	0.114	0.092	3	Do	0.085	-0.208
15	<i>Chlamydia pneumoniae</i> AR39	41	-0.034	0.137	-0.013	0.109	0.089	3	Do	0.075	-0.274
16	<i>Synechocystis</i> spPCC6803	48	0.558	-0.078	0.2	0.001	0.004	2	Cyanobacteria	0.514	-0.556
17	<i>Nostoc</i> sp	41	0.305	-0.042	-0.03	0.002	0.005	3	Cyanobacteria	0.299	-0.238
18	<i>Thermosynechococcus</i> <i>elongatus</i> BP-1	54	0.111	-0.14	0.066	0.002	0.014	3	Do	0.164	-0.432
19	<i>Deinococcus radiodurans</i> R1chromosome I	67	-0.42	0.566	0.055	0.003	0.024	2	Deinococci	-0.227	-0.558
20	<i>Deinococcus radiodurans</i> R1chromosome II	66	0.188	-0.181	0.082	0.036	0.024	3	Do	-0.227	-0.558
21	<i>Lactococcus lactis</i> subsp <i>lactis</i> IL1403	35	0.632	0.049	-0.113	0.149	0.117	1	Firmicutes	-0.024	-0.6
22	<i>Mycoplasma gallisepticum</i> strain R	31	-0.602	0.49	-0.167	0.054		1	Do	0.614	0.411
23	<i>Streptococcus agalactiae</i> 2603VR	36	-0.618	-0.234	-0.029	0.158	0.133	1	Do	0.037	-0.596
24	<i>Streptococcus pyogenes</i> MGAS2096	39	-0.64	0.238	-0.503	0.143	0.125	1	Do	0.246	-0.474
25	<i>Bacillus anthracis</i> Ames	35	0.548	0.527	0.276	0.271	0.174	2	Do	0.179	-0.561
26	<i>Streptococcus mutans</i> UA159	37	0.469	-0.011	0.217	0.12	0.111	2	Do	0.155	-0.242
27	<i>Oceanobacillus iheyensis</i> HTE831	36	0.445	-0.049	-0.027	0.141	0.088	2	Do	0.102	-0.292
28	<i>Streptococcus pneumoniae</i> R6	40	0.425	0.051	0.091	0.145	0.057	2	Do	0.209	-0.525
29	<i>Mycoplasma genitalium</i> G- 37	32	-0.47	0.096	0.195	0.027		2	Do	0.424	0.216
30	<i>Listeria monocytogenes</i> F6854	37	-0.52	0.211	-0.097	0.133	0.126	2	Do	0.037	-0.497
31	<i>Staphylococcus aureus</i> N315	33	-0.54	-0.046	-0.01	0.205	0.137	2	Do	0.039	-0.493
32	<i>Lactobacillus plantarum</i> WCFS1	44	0.339	0.413	-0.356	0.212	0.131	3	Do	-0.004	-0.653

A Statistical study on the nucleotide composition of bacterial chromosomes.

33	<i>Clostridium perfringens</i> 13	29	0.295	-0.002	-0.031	0.271	0.209	3	Do	0.096	-0.246
34	<i>Bacillus subtilis</i>	44	0.285	0.58	-0.192	0.091	0.089	3	Do	0.272	-0.388
35	<i>Mycoplasma penetrans</i> HF-2	26	0.236	-0.083	0.323	0.136	0.05	3	Do	0.413	0.307
36	<i>Bacillus halodurans</i> C-125	43	0.199	0.388	-0.076	0.102	0.094	3	Do	0.117	-0.785
37	<i>Mycoplasma pneumoniae</i>	40	0.114	0.253	-0.072	0.012	0.022	3	Do	0.027	-0.303
38	<i>Mycoplasma pulmonis</i> UAB CTIP	27	-0.084	-0.158	-0.345	0.005	0.041	3	Do	0.378	-0.033
39	<i>Clostridium tetani</i> E88	29	-0.112	0.057	-0.139	0.223	0.186	3	Do	0.056	0.022
40	<i>Clostridium acetobutylicum</i> ATCC824	39	-0.135	0.039	-0.272	0.332	0.19	3	Do	-0.01	-0.169
41	<i>Thermoanaerobacter tengcongensis</i> MB4(T)	38	-0.258	-0.157	-0.007	0.187	0.141	3	Do	0.217	-0.004
42	<i>Enterococcus faecalis</i> V583	38	-0.35	0.38	-0.415	0.13	0.112	3	Do	0.0461	-0.615
43	<i>Staphylococcus epidermidis</i> RP62A	32	-0.366	-0.054	0.236	0.129	0.115	3	Do	0.281	-0.107
44	<i>Fusobacterium nucleatum</i> ATCC 25586	27	0.118	-0.183	0.152	0.131	0.134	3	Fusobacteria	0.315	0.104
45	<i>Borrelia burgdorferi</i> B31	28	0.361	0.202	-0.022	0.155	0.213	3	Spirochaetes	0.039	-0.029
46	<i>Treponema pallidum</i> Nichols	53	0.057	0.035	0.32			3	Do	-0.105	-0.457
47	<i>Mesorhizobium loti</i> MAFF303099	63	0.678	0.003	-0.353	0.026	0.012	1	α -Proteobact	-0.483	-0.725
48	<i>Caulobacter crescentus</i> CB15	67	0.624	-0.677	0.209	0.035	0.025	1	Do	-0.527	-0.726
49	<i>Agrobacterium tumefaciens</i> C58 UWashCircular	59	-0.642	-0.612	-0.015	0.041	0.048	1	Do	-0.248	-0.559
50	<i>Brucella melitensis</i> 16M chr 1	57	0.549	-0.457	-0.13	0.064	0.055	2	Do	0.024	-0.67
51	<i>Sinorhizobium meliloti</i> 1021	63	-0.459	0.274	0.227	0.028	0.026	2	Do	-0.248	-0.559
52	<i>Bradyrhizobium japonicum</i> USDA 110	64	-0.58	0.02	-0.325	0.019	0.019	2	Do	-0.487	-0.63
53	<i>Rickettsia conorii</i> Malish 7	32	0.029	0.06	-0.079	0.097	0.05	3	Do	0.122	-0.091
54	<i>Rickettsia prowazekii</i>	29	-0.145	0.002	-0.032	0.128	0.07	3	Do	0.112	-0.131

A Statistical study on the nucleotide composition of bacterial chromosomes.

Madrid E											
55	<i>Ralstonia solanacearum</i> GMI1000	67	0.635	-0.34	-0.652	0.024	0.04	1	β-Proteobact	-0.455	-0.723
56	<i>Nitrosomonas europaea</i> ATCC 19718	51	-0.43	-0.088	0.325	0.095	0.057	2	Do	0.321	-0.611
57	<i>Neisseria meningitidis</i> FAM18	52	-0.09	-0.058	-0.711	0.081	0.06	3	Do	0.008	-0.299
58	<i>Escherichia coli</i> MG1655	51	0.832	0.558	-0.031	0.041	0.045	1	γ-Proteobact	0.385	-0.888
59	<i>Haemophilus influenzae</i> Rd KW20	38	0.764	0.007	0.173	0.069	0.107	1	Do	0.173	-0.646
60	<i>Pasteurella multocida</i> PM70	40	0.706	0.379	0.233	0.116	0.102	1	Do	0.094	-0.776
61	<i>Salmonella enterica</i> serovar Typhi CT18	52	0.685	-0.734	-0.372	0.068	0.062	1	Do	0.367	-0.813
62	<i>Xanthomonas axonopodis</i> pv. citri 306	65	-0.609	-0.165	0.148	0.035	0.029	1	Do	-0.392	-0.778
63	<i>Xanthomonas campestris</i> pv. vesicatoria str	65	-0.631	-0.177	0.084		0.029	1	Do	-0.418	-0.809
64	<i>Shewanella oneidensis</i> MR- 1	46	-0.775	-0.385	-0.433		0.082	1	Do	0.21	-0.743
65	<i>Pseudomonas syringae</i> DC3000	58	-0.807	0.369	0.035		0.049	1	Do	0.455	-0.837
66	<i>Pseudomonas putida</i> F1	61	-0.836	-0.343	0.081	0.032	0.04	1	Do	-0.28	-0.857
67	<i>Vibrio cholerae</i> O395 chromosome 1	47	0.563	0.53	0.7	0.075	0.064	2	Do	0.354	-0.785
68	<i>Pseudomonas aeruginosa</i> PAO1	67	-0.51	0.55	0.293	0.055	0.063	2	Do	-0.338	-0.62
69	<i>Buchnera aphidicola</i> (Baizongia pistaciae)	25	0.046	-0.05	-0.038	0.315	0.267	3	Do	0.138	-0.073
70	<i>Coxiella burnetii</i> RSA	43	0.011	-0.05	0.094		0.035	3	Do	-0.139	-0.321
71	<i>Xylella fastidiosa</i> 9a5c	53	-0.154	0.024	0.07	0.142	0.116	3	Do	0.16	-0.038
72	<i>Buchnera</i> sp APS	26	-0.227	-0.024	-0.137		0.074	3	Do	-0.106	-0.196
73	<i>Buchnera aphidicola</i> Sg	25	-0.24	-0.275	0.362	0.053	0.057	3	Do	-0.087	-0.148
74	<i>Wigglesworthia glossinidia</i>	22	-0.365	0.352	0.344		0.073	3	Do	-0.009	-0.255

A Statistical study on the nucleotide composition of bacterial chromosomes.

	brevipalpis										
75	<i>Campylobacter jejuni</i> NCTC 11168	31	-0.177	-0.013	-0.404	0.193	0.068	3	ε-Proteobact	0.372	0.237
76	<i>Helicobacter pylori</i> J99	39	-0.246	0.125	0.115	0.075	0.056	3	do	0.095	-0.529

r_1 , r_2 and r_3 are the Pearson correlations of UCU(g) values with 1st, 2nd and 3rd axis coordinates generated by correspondence analysis. mut_{ir} and mut_{ir} are the strand specific mutational bias in the intergenic region and synonymous third codon position respectively (Materials and Methods). Third position skew in coding sequences we could not calculate (grey boxes) in 9 strains as it was confusing in identifying the genes to their strand locations from the locus tag. In one of the genomes *Treponema pallidum* Nichols, the intergenic skew could not be calculated (grey boxes) because of the unavailability of the sequences belonging to this strain.

Bacteria analyzed in this study have been divided into three categories 1, 2 and 3.

Category 1: high correlation ($|r| \geq 0.6$) between UCU(g) and principal axis; Category 2: weak correlation ($0.6 > |r| \geq 0.4$) with the principal axis; Category 3: very weak correlation ($|r| < 0.4$) with the primary axis.

4.5. Discussion

According to the selection-mutation-drift theory (Sharp and Li, 1986; Bulmer, 1991) selected codon usage bias is correlated with gene expression. The positive correlation of UCU(g) with CAI (Sharp and Li, 1987b) values as well as with expressions found in this study suggests that UCU(g) is a potential measure of the selected codon usage bias in genes of *E. coli* and *S. cerevisiae*, two organisms belonging to different kingdoms of life. Out of 20 bacterial genomes exhibiting strong correlation between UCU(g) and the primary axis of correspondence analysis, 13 have been shown to have strong selected codon usage bias having 'S' value ≥ 1.0 (Sharp *et al*, 2005). The rest 7 bacterial genomes have the 'S' ≥ 0.5 except *R. solanacearum*. The correlation result of UCU(g) with primary axis of correspondence analysis in different bacteria known to have strong selected codon usage bias indicates the applicability of UCU(g) in other bacterial genomes. The strong correlation between UCU(g) and ENc in several bacteria is also in support of this statement. The lack of

correlation result between UCU(g) and ENc in case of *Treponema pallidum*, *X. fastidiosa* etc also favours that UCU(g) measures the selected codon usage bias in genes.

Sharp and co-workers observed a variation in the strength of selected codon usage bias 'S' among bacteria (Sharp *et al.*, 2005). 'S' is calculated from the codon usage of phenylalanine, tyrosine, isoleucine and asparagines. We compared our result with the result of Sharp *et al.* (2005) in a systematic manner. Out of the 8 Actinobacteria genomes analysed in this study, only *T. whipplei* Twist was to have weak selected codon usage bias which is in concordant with 'S' value. In addition we found weak selected codon usage bias in *S. coelicolor*, *S. avermitilis* which is unlike the 'S' value. These bacteria are having GC rich genomes and the high 'S' values associated with these genomes might be driven by mutational bias. This is explained in Sharp *et al.* study because randomized value is very much equivalent to 'S' value. In addition, we have observed correlation between UCU(g) and ENc. So, selected codon usage bias is weak in these genomes. Our result was found different in case of Firmicutes from that of Sharp *et al.* (2005). Out of 23 bacterial genomes only seven genomes were observed with high selected codon usage bias by our approach. In contrast, except Mollicutes (*Mycoplasma* and *Ureaplasma*) and *T. tengcongensis*, all other Firmicute genomes were reported to have strong selected codon usage bias (Sharp *et al.*, 2005). Both the *Clostridium* genomes were found with low selected codon usage bias by our approach, but these were shown with high selected codon usage bias by Sharp *et al.* (2005). The different result in case of Firmicutes indicates UCU(g) may not be efficient in case of Firmicutes, which are AT rich genomes. This was found out to be true after studying the codon usage in *C. perfringens*. In AT rich genomes, the family box codons CGN (arg) and CTN (leu) are used significantly less. These genomes use synonymous codons AGR (arg) and TTR (leu). Due to lesser use of the family box codons, UCU(g) measure may not correctly reflect the selection. As the number of family box decreases the chance of variation among the RSCU values ending with the same nucleotide also decreases upon which our method is dependent. UCU(g) measure based on eight family boxes is always better than the measure based on five family box codons as observed in *E. coli* genome. The correlation value between UCU(g) and expression decreases when the former is calculated from five family box codons in stead of eight family box codons.

We observed that the selected codon usage bias is more in *P. Aeruginosa*, which is in concordant with the report published earlier (Grocock and Sharp, 2002). But according to Sharp *et al* (2005) selected codon usage bias is low in this bacterium. Sharp *et al* (2005) have already explained in their paper that the four amino acid codons used for their study do not show significant selection but selection is found more in other amino acid codons. So, with reference to the genomes of *P. aeruginosa*, *Streptomyces coelicolor* and *Streptomyces avermitilis*, application of UCU(g) method seems a better approach than that of Sharp *et al*'s (2005). These genome are GC rich, which implies that UCU(g) application gives better result in these genomes.

Apart from the application in AT rich genomes, the other demerit of UCU(g) is disentangling the effect of selection and mutation where the effect of both the factors acts in the same direction (Li, 1987; McVean and Charlesworth, 1999). The example follow is from the Sharp *et al* (2005) paper. In phenylalanine, tyrosine, isoleucine and asparagines, always the 'C' ending codons are selected over the U ending codons in the highly expressed genes for efficient translation. In GC rich genomes, the same preference is due to mutation and not due to selection. So it is difficult to separate the individual effects in this case as seen in the case of *Streptomyces coelicolor* and *Streptomyces avermitilis*. This indicates that it is difficult to find out a single approach that will accurately measure the selected codon usage bias in a gene.

Apart from the genome GC%, the mutational biases between the strands (Frank and Lobry, 1999) and along a replicore in chromosomes (Daubin and Perriere, 2003; Arakawa and Tomita, 2007) are well established in bacteria. The correlation is very strong between the strand specific mutational bias at the intergenic region and the same at the 3rd position of codons (this study; Lobry and Sueoka, 2002), which suggests that mutational bias affects significantly codon usage in genes. However, it is observed in this study that the effect of strand specific mutational bias on selected codon usage bias is negligible. This suggests that the effect of selection forces over mutational bias is dominant during evolution.

Conclusion

5. Conclusion

The compositional analysis of bacterial chromosomes seems to be interlinked with respect to the three phenomena i.e. Chargaff's second parity, strand specific mutational bias and selected codon usage bias. In the literatures that we have reviewed in Chapter I, these three phenomena have been found to be addressed from different angles. Although a large number of literature and experiments are found but these are still not exhaustive.

Two new methodologies have been developed in this study. The first one is named Intra-strand frequency distribution parity, which has been used to study PR2 in entire chromosomes. Using this methodology it has been observed that violation of PR2 is observed in several bacterial chromosomes. Strand specific mutational bias, asymmetry in replication topography and gene compositional asymmetry between the strands were found affecting PR2 in chromosomes. However, no single reason was found to be sufficient enough to explain the violation of parity. Bacteria with high SSMB and high asymmetry in replication topography are likely to violate the PR2 in chromosomes as we observed in the case of *Xylella fastidiosa*. In addition gene distribution asymmetry between the strands will also contribute to the PR2 violation. Both gene distribution asymmetry and SSMB are high in Firmicutes. This is the reason for the two Firmicute genomes having high asymmetry in replication topography possesses the GCS > 0.01 and violates parity (Table 2.1). Genomes with high GCS were all shown to violate ISFDP, but genomes with low GCS were also found to violate PR2, which was surprising. This indicates that our understanding of different causative factors for ISFDP violation in genomes is incomplete.

The second new method developed in this study is "unevenness of codon usage". The underlying hypothesis for developing this method comes from our understanding about the role of tRNA as a selective factor for codon usage bias. This is an indirect approach to measure the amount of selection responsible of the codon usage bias in a gene. The methodology has been shown to work significantly against the proteome data now available for *E. coli* and *S. cerevisiae*. Results obtained in different bacterial genomes with the new codon usage measure are found encouraging with some limitations in AT rich genomes. The other demerit of this approach is to disentangle the mutation and selection when both act in

the same direction. However, we feel the method will be applicable in several genomes as we have already demonstrated in Chapter IV. This measure might be of interest to molecular evolutionary biologists as it is not dependent upon the gene expression data for finding the selected codon usage bias in genes.

The recent available proteome data of *E. coli* was used to study the effect of strand specific mutational bias, which was found to be less effective against the highly expressed genes than weakly expressed genes in *E. coli*. In relation to codon usage analysis as well as strand specific mutational bias study we observed that strand specific mutational bias plays a negligible role in determining codon usage in highly expressed gene in bacteria with high selected codon usage bias.

There are several issues yet to be addressed in genome compositional analysis. The role of selection mechanism i.e. DNA structure in maintaining PR2 in chromosomes is an important aspect as reviewed by Forsdyke and Mortimer (2000). Bacterial chromosomes vary with respect to strand specific mutational bias. In Firmicutes the mutational bias is observed very high which coincides with the higher gene compositional asymmetry between the strands. Does the secondary structures in the chromosomes of Firmicutes are less formed in comparison to other group of bacteria? This is yet to be known. The higher skew at 3rd position of codons than at the intergenic region is yet an unsolved problem. If cytosine deamination is one of the main reason for the observation of skew in genomes then why in Firmicutes 'A' is found more than 'T' in leading strands?

An interesting aspect of codon usage bias is observed in human genome where the nucleotide composition plays a dominant role in determining codon usage. Whether selection plays a role in determining codon usage bias or it is driven by mutational bias is yet to be understood in human. The future objectives of this study will be to find out the reason for strand specific mutational bias in bacteria as well as codon usage bias in human.

6. References:

1. Adams, M.J. and Antoniw, J.F. DPVweb: An open access internet resource on plant viruses and virus diseases, *Outlooks on Pest Management* **16**, 268-270 (2005).
2. Adams, M.J. and Antoniw, J.F. DPVweb: a comprehensive database of plant and fungal virus genes and genomes, *Nucleic Acids Res.* **34**, Database issue, D382-D385. (<http://www.dpvweb.net/>) (2006).
3. Albrecht-Buehler, G. Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions, *Proc. Natl. Acad. Sci., USA* **103**, 17828-17833 (2006).
4. Arakawa, K. and Tomita, M. Selection effects on the positioning of genes and gene structures from the interplay of replication and transcription in bacterial genomes, *Evolutionary Bioinformatics* **3**, 279-286 (2007).
5. Baisnée, P.F., Hampson, S. and Baldi, P. Why are complementary DNA strands symmetric? *Bioinformatics* **18**, 1021-1033 (2002).
6. Bell, S.J. and Forsdyke, D.R. Accounting units in DNA, *J. Theor. Biol.* **197**, 51-61 (1999).
7. Benzécri, J. L'Analyse des Données. Volume II. L'Analyse des Correspondances. Paris, France: Dunod. (1973).
8. Bennetzen, J.L. and Hall, B.D. Codon selection in yeast, *J. Biol. Chem.* **257**, 3026-3031 (1982).
9. Bernardi, G. and Bernardi, G. Compositional constraints and genome evolution. *J. Mol. Evol.* **24**, 1-11 (1986).
10. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. The mosaic genome of warm-blooded vertebrates, *Science* **228**, 953-958 (1985).

11. Bremer, H. and Dennis, P.P. Modulation of chemical composition and other parameters of the cell by growth rate in *Escherichia coli* and *Salmonella*, *Cellular and Molecular Biology* (eds F. Neidhardt and others), pp. 1553–1569. Washington DC: American Society for Microbiology (1996).
12. Bulmer, M. Coevolution of codon usage and transfer RNA abundance, *Nature* **325**, 728-730 (1987).
13. Bulmer, M. Are codon usage patterns in unicellular organisms determined by a mutation-selection balance? *J. Evol. Biol.* **1**, 15–26 (1988).
14. Bulmer, M. The selection-mutation-drift theory of synonymous codon usage, *Genetics* **129**, 897-907 (1991).
15. Chargaff, E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation, *Experientia* **6**, 201 -209 (1950).
16. Chargaff, E. Structure and function of nucleic acids as cell constituents, *Fed. Proc.* **10**, 654-659. (1951).
17. Chen, L. and Zhao, H. Negative correlation between compositional symmetries and local recombination rates, *Bioinformatics* **21**, 3951-3958 (2005).
18. Chen, S.L., Lee, W., Hotts, A.K., Shapiro, L. and McAdams, H.H. Codon usage between genomes is constrained by genome wide mutational processes, *Proc. Natl. Acad. Sci. USA* **101**, 3480–3485 (2004).
19. Coghlan, A. and Wolfe, K.H. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*, *Yeast* **16**, 1131–1145, (2000).
20. Daubin, V. and Perriere, G. G+C structuring along the genome: a common feature in prokaryotes, *Mol. Biol. Evol.* **20**, 471–483 (2003).
21. Deng, B. Mismatch repair error implies Chargaff's second parity rule, arXiv:0704.2191v2[q-bio.GN],1-9
[http://arxiv.org/PS_cache/arxiv/pdf/0704/0704.2191v2.pdf] (2007).

-
22. dos Reis M., Wernisch, L. and Savva, R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome, *Nucleic Acids Res.* **31** 6976–6985 (2003).
 23. dos Reis, M., Savva, R., and Wernisch, L. Solving the riddle of codon usage preferences, a test for translational selection, *Nucleic Acids Res.* **32**, 5036–5044 (2004).
 24. dos Reis, M. and Wernisch, L. Estimating translational selection in eukaryotic genomes, *Mol. Biol. Evol.* **26**, 451–461 (2009).
 25. Duret, L. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes, *Trends Genet* **16**, 287-289 (2000).
 26. Duret, L. and Mouchiroud, D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*, *Proc. Natl. Acad. Sci. USA* **96**, 4482–4487 (1999).
 27. Duret, L., Eyre-Walker, A. and Galtier, N. A new perspective on isochore evolution, *Gene* **385**, 71–74 (2006).
 28. Ermolaeva, M.D. Synonymous codon usage in bacteria, *Curr. Issues Mol. Biol.* **3**, 91–97 (2001).
 29. Eyre-Walker, A. and Bulmer, M. Synonymous substitution rates in enterobacteria, *Genetics* **140**, 1407–1412 (1995).
 30. Fijalkowska, I.J., Jonczyk, P., Tkaczyk, M.M., Bialoskorska, M. and Schaaper, R.M. Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome, *Proc. Natl. Acad. Sci. USA* **95**, 10020-10025 (1998).
 31. Forsdyke, D.R. A stem-loop ‘kissing’ model for the initiation of recombination and the origin of introns, *Mol. Biol. Evol.*, **12**, 949–958 (1995).
 32. Forsdyke, D.R. and Mortimer, J.R. Chargaff’s legacy, *Gene* **261**, 127-137 (2000).

-
33. Francino, M.P., Chao, L., Riley, M.A. and Ochman, H. Asymmetries generated by transcription-coupled repair in enterobacterial genes, *Science* **272**, 107–109 (1996).
 34. Francino, M.P. and Ochman, H. Strand asymmetries in DNA evolution, *Trends Genet.* **13**, 240–245 (1997).
 35. Francino, M.P. and Ochman, H. Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences, *Mol. Biol. Evol.* **18**, 1147-1150 (2001).
 36. Frank, A.C. and Lobry, J.R. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms, *Gene* **238**, 65-77 (1999).
 37. Frank, A.C. and Lobry, J.R. Oriloc: prediction of replication boundaries in annotated bacterial chromosomes, *Bioinformatics* **16**, 560-561 (2000).
 38. Freeman, J.M., Plasterer, T.N., Smith, T.F. and Mohr, S.C. Patterns of genome organization in bacteria, *Science* **279**, 1827a (1998).
 39. Freire-Picos, M.A., Gonazalez-Siso, M.I., Rodriguez-Belmonte, E., Rodriguez-Torres, A.M., Ramil, E. and Cerdan, M.E. Codon usage in *Kluyveromyces lactis* and yeast cytochrome *c*-ending genes, *Gene* **139**, 43-49 (1994).
 40. Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. Global analysis of protein expression in yeast, *Nature* **425**, 737–741 (2003).
 41. Gautier, C. Compositional biases in DNA, *Current Opinion in Genetics and Development*, **10**, 656-661 (2000)
 42. Gouy, M. and Gautier, C. Codon usage in bacteria: correlation with gene expressivity, *Nucleic Acids Res.* **10**, 7055–74 (1982).
 43. Grantham, R., Gautier, C. and Gouy, M. Codon frequencies in 119 genes confirm consistent choices of degenerate base according to genome type, *Nucleic Acids Res.* **8**, 1892-1912 (1980a).
 44. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A. Codon catalogue usage and the genome hypothesis, *Nucleic Acids Res.* **8**, r49-r62 (1980b).

-
45. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. Codon catalogue usage is a genome strategy for genome expressivity, *Nucleic Acids Res.* **9**, r43-r75 (1981).
 46. Green, P., Ewing, B., Miller, W., Thomas, P.J. and Green, E.D. NISC Comparative Sequencing Program Transcription-associated mutational asymmetry in mammalian evolution, *Nature Genet* **33**, 514–517 (2003).
 47. Grigoriev, A. Analyzing genomes with cumulative skew diagrams, *Nucleic Acid Res.* **26**, 2286-2290 (1998).
 48. Grocock, R.J. and Sharp, P.M. Synonymous codon usage in *Pseudomonas aeruginosa* PAO1, *Gene* **289**, 131-139 (2002).
 49. Grosjean, H. and Fiers, W. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes, *Gene* **18**, 199–209 (1982).
 50. Guindon, S. and Perriere, G. Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes, *Mol. Biol. Evol.* **18**, 1838 – 1840 (2001).
 51. Hendrickson, H. and Lawrence, J.G. Selection for chromosome architecture in bacteria, *J. Mol. Evol.* **62**, 615-629 (2006).
 52. Henry, I. and Sharp, P.M. Predicting gene expression level from codon usage bias, *Mol. Biol. Evol.* **24**, 10–12, (2007).
 53. Hershberg, R. and Petrov, D.A. Selection on codon bias, *Annu. Rev. Genet.* **42**, 287–299 (2008).
 54. Higgs, P.G. and Ran, W. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage, *Mol. Biol. Evol.* **25**, 2279-2291 (2008).
 55. Hinds, P.W. and Blake, R.D. Degrees of divergence in the *E. coli* genome from correlations between dinucleotide, trinucleotide and codon frequencies, *J. Biomol. Struct. Dyn.* **2**, 101-118 (1984).

-
56. Hinds, P.W. and Blake, R.D. Delineation of coding areas in DNA sequences through assignment of codon probabilities, *J. Biomol. Struct. Dyn.* **3**, 543–549 (1985).
 57. Hiraoka, Y., Kawamata, K., Haraguchi, T. and Chikashige, Y. Codon usage bias is correlated with gene expression levels in the fission yeast *Schizosaccharomyces pombe*, *Genes to Cells* **14**, 499–509 (2009).
 58. Hu, J., Zhao, X. and Yu, J. Replication-associated purine asymmetry may contribute to strand-biased gene distribution, *Genomics* **90**, 186–194 (2007).
 59. Ikemura, T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes, a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, *J. Mol. Biol.* **151**, 389–409 (1981).
 60. Ikemura, T. Correlation between the abundance of yeast tRNAs and the occurrence of the respective codons in its protein genes, *J. Mol. Biol.* **158**, 573–597 (1982).
 61. Ikemura, T. Codon usage and transfer-RNA content in unicellular and multicellular organisms, *Mol. Biol. Evol.* **2**, 13–34 (1985).
 62. Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F.U., Kerner, M.J. and Frishman, D. Protein abundance profiling of the *Escherichia coli* cytosol, *BMC Genomics* **9**, 102 (2008).
 63. Johnson, A. and O'Donnell, M. Cellular DNA replicases: components and dynamics at the replication fork, *Annu. Rev. Biochem.* **74**, 283–314 (2005).
 64. Kano-Sueoka, T., Lobry, J.R. and Sueoka, N. Intra-strand biases in bacteriophage T4 genome, *Gene* **238**, 59–64 (1999).
 65. Konigsberg, W.J.N. and Godson, G.N. Evidence for use of rare codons in the *dnaG* gene and other regulatory genes of *Escherichia coli*, *Proc. Natl. Acad. Sci. USA* **80**, 687–691 (1983).
 66. Li, W.H. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons, *J. Mol. Evol.* **24**, 337–345 (1987).

-
67. Lobachev, K.S., Shor, B.M., Tran, H.T., Taylor, W., Keen, J.D., Resnick, M.A. and Gordenin, D.A. Factors affecting inverted repeat stimulation of recombination and deletion in *Saccharomyces cerevisiae*, *Genetics* **148**, 1507-1524 (1998).

 68. Lobry, J.R. and Sueoka, N. Asymmetric directional mutation pressures in bacteria, *Genome Biology* **3**, research0058.1-0058.14 (2002).
 69. Lobry, J.R. Asymmetric substitution patterns in the two DNA strands of bacteria, *Mol. Biol. Evol.* **13**, 660–665 (1996).
 70. Lobry, J.R. Properties of a general model of DNA evolution under no-strand bias conditions, *J. Mol. Evol.* **40**, 326-330 (1995).

 71. Lowe, T.M. and Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genome sequences, *Nucleic Acid Res.* **25**, 955–964 (1997).

 72. Mackiewicz, P., Gierlik, A., Kowalczyk, M., Dudek, M.R. and Cebrat, S. How does replication-associated mutational pressure influence amino acid composition of proteins? *Genome Res.* **9**, 409–416 (1999).

 73. McInerney, J.O. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*, *Proc. Natl. Acad. Sci. USA* **95**, 10698–10703 (1998).

 74. Mclean, M.J., Wolfe, K.H. and Devine, K.M. Base composition skews, replication orientation and gene orientation in 12 prokaryote genomes, *J. Mol. Evol.* **47**, 691–696 (1998).

 75. McVean, G.A.T. and Charlesworth, B. A population genetic model for the evolution of synonymous codon usage: patterns and predictions, *Genet. Res.* **74**, 145–158 (1999).

 76. Mira, A. and Ochman, H. Gene Location and Bacterial Sequence Divergence, *Mol. Biol. Evol.* **19**, 1350-1358 (2002).

 77. Mitchell, D. and Bridge, R. A test for Chargaff's second rule, *Biochem. Biophys. Res. Comm.* **340**, 90-94 (2006).

-
78. Morton, R.A. and Morton, B.R. Separating the effects of mutation and selection in producing DNA skew in bacterial chromosomes, *BMC Genomics* **8**, 368 (2007).
 79. Mrazek, J. and Karlin, S. Strand compositional asymmetry in bacterial and large viral genomes, *Proc. Natl. Acad. Sci. USA* **95**, 3720-3725 (1998).
 80. Muto, A. and Osawa, S. The guanine and cytosine content of genomic DNA and bacterial evolution, *Proc. Natl. Acad. Sci. USA* **84**, 166–169 (1987).
 81. Neçşulea, A. and Lobry, J.R. A new method for assessing the effect of replication on DNA base composition asymmetry, *Mol. Biol. Evol.* **24**, 2169–2179 (2007).
 82. Nikolaou, C. and Almirantis, Y. A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species, *Nucleic Acid Res.* **33**, 6816-6822 (2005).
 83. Nikolaou, C. and Almirantis, Y. Deviations from Chargaff's second parity rule in organellar DNA Insights into the evolution of organellar genomes, *Gene* **381**, 34–41 (2006).
 84. Nussinov, R. Some indications for inverse DNA duplication, *J. Theor. Biol.* **95**, 783–793 (1982).
 85. Nussinov, R. Strong doublet preferences in nucleotide sequences and DNA geometry, *J. Mol. Evol.* **20**, 111-119 (1984).
 86. Ochman, H. Neutral mutations and neutral substitutions in bacterial genomes, *Mol. Biol. Evol.* **20**, 2091–2096 (2003).
 87. Okamura, K., Wei, J. and Scherer, S.W. Evolutionary implications of inversions that have caused intra-strand parity in DNA, *BMC Genomics* **8**, 160 (2007).
 88. Osawa, S., Jukes T.H., Watanabe, K. and Muto, A. Recent evidence for evolution of the genetic code, *Microbiol. Rev.* **56**, 229-264 (1992).
 89. Peden, J.F. CodonW. PhD Thesis, University of Nottingham (1999).

-
90. Poptsova, M.S., Larionov, S.A., Ryadchenko, E.V. Rybalko, S.D., Zakharov, I.A. and Loskutov, A. Hidden chromosome symmetry: *in silico* transformation reveals symmetry in 2D DNA walk trajectories of 671 chromosomes, *PLoS ONE* **4**, e6396 (2009).
 91. Powdel, B.R., Satapathy, S.S., Kumar, A, Jha, P.K., Buragohain, A.K., Borah, M. and Ray, S.K. A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (Chargaff's Second Parity Rule). *DNA Res.* **16**, 325–343 (2009).
 92. Powdel, B.R., Borah, M. and Ray, S.K. Strand-specific mutational bias influences codon usage of weakly expressed genes in *Escherichia coli*, *Genes to Cells* **15**, 773-782 (2010).
 93. Prabhu, V.V. Symmetry observed in long nucleotide sequences, *Nucleic Acid Res.* **21**, 2797-2800 (1993).
 94. Qi, D. and Cuticchia, A.J. Compositional symmetries in complete genomes, *Bioinformatics* **17**, 557-559 (2001).
 95. Rachidi, N., Martinez, M.J., Barre, P. and Blondin, B. *Saccharomyces cerevisiae* PAU genes are induced by anaerobiosis, *Mol. Microbiol.* **35**, 1421–1430 (2000).
 96. Rocha, E.P.C., Danchin, A. and Viari, A. Universal replication biases in bacteria, *Mol. Microbiol.* **32**, 11–16 (1999).
 97. Rocha, E.P.C. and Danchin, A. Gene essentiality determines chromosome organization in bacteria, *Nucleic Acid Res.* **31**, 6570-6577 (2003).
 98. Rocha, E.P.C. The replication-related organization of bacterial genomes, *Microbiol.* **150**, 1609-1627 (2004).
 99. Rocha, E.P.C., Touchon, M. and Feil, E.J. Similar compositional biases are caused by very different mutational effects, *Genome Res.* **16**, 1537–1547 (2006).
 100. Rocha, E.P.C. The organization of the bacterial genome, *Annu. Rev. Genet.* **42**, 211-233 (2008).

-
101. Rudner, R., Karkas, J.D. and Chargaff, E. Separation of *B. subtilis* DNA into complementary strands, III. Direct analysis, *Proc. Natl. Acad. Sci. USA*, **60**,921-922 (1968).
 102. Rudner, R., Karkas, J.D. and Chargaff, E. Separation of microbial deoxyribonucleic acids into complementary strands, *Proc. Natl. Acad. Sci. USA*, **63**,152-159 (1969).
 103. Satapathy, S.S., Dutta, M. and Ray, S.K. Variable correlation of genome GC% with transfer RNA number as well as with transfer RNA diversity among bacterial groups: α -Proteobacteria and Tenericutes exhibit strong positive correlation. *Microbiol. Res.* **165**, 232–242 (2010).
 104. Schmid, M.B. and Roth, J.R. Gene location affects expression level in *Salmonella typhimurium*, *J. Bacteriol.* **169**, 2872-2875 (1987).
 105. Seğmon, M., Lobry, J.R. and Duret, L. No evidence for tissue-specific adaptation of synonymous codon usage in humans. *Mol. Biol. Evol.* **23**, 523–529 (2006).
 106. Sernova, N.V. and Gelfand, M.S. Identification of replication origins in prokaryotic genomes, *Brief. Bioinformatics* **9**, 376–91 (2008).
 107. Sharp, P.M. and Li, W.H. An evolutionary perspective on synonymous codon usage in unicellular organisms, *J. Mol. Evol.* **24**, 28–38 (1986a).
 108. Sharp, P.M. and Li, W.H. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons, *Nucleic Acids Res.* **14**, 7737–7749 (1986b).
 109. Sharp, P.M. and Li, W.H. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias, *Mol. Biol. Evol.* **4**, 222-230 (1987a).
 110. Sharp, P.M. and Li, W.H. The codon adaptation index – a measure of directional codon usage bias, and its potential application, *Nucleic Acid Res.* **15**, 1281-1295 (1987b).
 111. Sharp, P.M., Shields, D.C., Wolfe, K.H. and Li, W.H. Chromosomal location and evolutionary rate variation in enterobacterial genes, *Science* **246**, 808-810 (1989).
-

-
112. Sharp, P.M. Determination of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution, *J. Mol. Evol.* **33**, 23-33 (1991).
 113. Sharp, P.M., Averof, M., Lloyd, A.T., Matassi, G. and Peden, J.F. DNA sequence evolution, the sounds of silence, *Phil. Trans. R. Soc. Lond. B.* **349**, 241–247 (1995).
 114. Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F. and Sockett, R.E. Variation in the strength of selected codon usage bias among bacteria, *Nucleic Acids Res.* **33**, 1141–1153 (2005).
 115. Sharp, P.M., Tuohy, T.M. and Mosurski, K.R. Codon usage in yeast, cluster analysis clearly differentiates highly and lowly expressed genes, *Nucleic Acids Res.* **14**, 5125–5143 (1986).
 116. Shields, D.C., Sharp, P.M., Higgins, D.G. and Wright F. “Silent” sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons, *Mol. Biol. Evol.* **5**, 704-716 (1988).
 117. Shields, D.C. Switches in species-specific codon preferences: the influence of mutation biases, *J. Mol. Evol.* **31**, 71-80 (1990).
 118. Shioiri, C. and Takahata, N. Skew of mononucleotide frequencies, relative abundance of dinucleotides and DNA strand asymmetry, *J. Mol. Evol.* **53**, 364–76 (2001).
 119. Smithies, O., Engels, W.R., Devereux, J.R., Slightom, J.L. and Shen, S-H. Base substitutions, length differences and DNA strand asymmetries in the human γ and α fetal globin gene region, *Cell* **26**: 345-353 (1981).
 120. Sueoka, N. On the genetic basis of variation and heterogeneity of DNA base composition, *Proc. Natl. Acad. Sci. USA*, **48**, 582-592 (1962).

-
121. Sueoka, N. Intra-strand parity rules of DNA base composition and usage biases of synonymous codons, *J. Mol. Evol.* **40**, 318-325 (1995).
 122. Sueoka, N. Translation-coupled violation of parity rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position, *Gene* **238**, 53-58 (1999).
 123. Suzuki, H., Brown, C.J., Forney, L.J. and Top, E.M. Comparison of correspondence analysis methods for synonymous codon usage in bacteria. *DNA Res.* **15**, 357-365, (2008).
 124. Szybalski, W., Kubinski, H. and Sheldrick, P. Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis, Cold Spring Harbor Symp. *Quant. Biol.* **31**, 123-127 (1966).
 125. Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths, *PLoS Genet.* **5**, e1000344 (2009).
 126. Verma, S.K., Das, D., Satapathy, S.S., Buragohain, A.K. and Ray, S.K. Compositional Symmetry of DNA duplex in bacterial genomes, *Curr. Sci.* **89**, 374-384 (2005).
 127. Walker, J.E., Saraste, M. and Gay, N.J. The unc operon, nucleotide sequence, regulation and structure of ATP-synthase, *Biochim. Biophys. Acta* **768**, 164-200 (1984).
 128. Watson, J.D. and Crick, F.H.C. Molecular structure of nucleic acids: A structure for deoxyribonucleic acid, *Nature* **171**, 737-738 (1953).
 129. Woodside, M.T., Behnke-Parks, W.M., Larizadeh, K., Travers, K., Herschlag, D. and Block S.M. Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins, *Proc. Natl. Acad. Sci. USA* **103**, 6190-6195 (2006).

130. Worning, P., Jensen, L.J., Hallin, P.F., Staerfeldt, H-H. and Ussery, D.W. Origin of replication in circular prokaryotic chromosomes, *Environ. Microbiol.* **8**, 353–61 (2006).
131. Wright, F. The ‘effective number of codons’ used in a gene, *Gene* **87**, 23-29 (1990).
132. Wu, C.I. and Maeda, N. Inequality in mutation rates of the two strands of DNA, *Nature* **327**, 169-170 (1987).
133. Yang, Z. and Nielsen, R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage, *Mol. Biol. Evol.* **25**, 568–579 (2008).
134. Zeeberg, B. Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes, *Genome Res.* **12**, 944-955 (2002).

List of Publications**Manuscripts****Published**

1. Powdel, B.R., Borah, M. and Ray, S.K. Strand specific mutational bias influences codon usage of weakly expressed genes in *Escherichia coli*. *Genes to Cells* **15**, 773-782 (2010).
2. Powdel, B.R., Satapathy, S.S., Kumar, A, Jha, P.K., Buragohain, A.K., Borah, M. and Ray, S.K. A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (Chargaff's second parity rule). *DNA Res.* **16**, 325-343 (2009).

Under revision

1. Satapathy, S.S., Powdel, B.R., Borah, M, Dutta, M., Ray, S.K. A new approach to study unevenness of codon usage in organisms without using a predefined gene set of highly expressed genes.