Computational Study of Origin of

Replication and ARS Sequences in Few Genomes

A thesis submitted in partial fulfillment of the requirements
for the award of the degree of

Master of Technology
In
Computational and Systems Biology

Submitted By:

Sanjeev Kumar
Enrol. No. 10/75/MT/012

Under the Supervision of
Dr. A. Krishnamachari

Centre for Computational Biology and Bioinformatics
School of Computational and Integrative Sciences
Jawaharlal Nehru University
New Delhi – 110067
(2010-2012)

I would like to dedicate this thesis to my loving and always supportive

Parents and family members …….

Acknowledgements

I take this opportunity to express a deep sense of gratitude towards my supervisor **Dr. A. Krishnamachari** for providing excellent guidance and encouragement throughout the project work. He introduced me to this area of work and was always there to help me in case of any need. Without his constant support and motivation this work could not have been completed successfully.

My sincere thanks to our dean, Professor Karmeshu. Also I would like to thank Prof. Indira Ghosh. She provided us with all the help and ensured that we do not have to face any problem in getting all the resources required for carrying out the research work.

I gratefully acknowledge my teachers Prof. Alok Bhattacharya, Prof. Ram Ramaswamy, Dr. N. Subba Rao, Dr. T.V. Vijay Kumar, Dr. Kushal Shah, Dr. Andrew Lynn, Dr. Pradipta Bandopadhyay, Dr. Lovekesh Vig, Dr. Rashi Gupta and Dr. Narinder Singh Sahni for providing their valuable guidance throughout the course.

I also want to acknowledge all classmates to provide me their help and valuable suggestions. My Special thanks to Mr. Tanmaya Meher and Mr. Vinod kumar Singh (Ph.D Research Scholor SC&IS,JNU) who motivated and discussed various matters with me during my project. I would also like to thank the staff of the school for their support and help.

Words fail me to express my appreciation to my family and friends for their constant support and motivation which was indispensable for successful completion of the project.

I acknowledge the financial assistance and resources provided by D.B.T. and J.N.U.

Finally, I would like to thank everybody who contributed to the successful completion of the work in any way and express my apology to those whose names I could not mention.

Sanjeev Kumar
M.Tech IInd YEAR
SCIS,JNU

Contents

List of Figures

List of Tables

CHAPTER-1

Introduction

1.1 Motivation

One of the fundamental processes happening in the cell cycle is the replication of DNA and a growing new cell must copy the genomic DNA before the cell division. DNA replication is a very complex process which includes the selection of initiation sites, unwinding of the DNA helix and assembly of the replication machinery. Replication may proceed bidirectional or unidirectional. Due to its central role in the cell cycle, identification of the origin of replication in various organisms is also important for discovery and development of new drugs for treatment of various diseases [8] .

The Autonomous Replicating Sequence [ARS element] is generally consists of A1, B1, B2 and B3 element. ARS constitutes eleven (11) bp sequence [(A/T)TTTAT(A/G)TTT(A/T)] is referred to as ARS consensus sequence i.e. (ACS) . Origin recognition complex binds to this consensus sequence A1 and initiates the replication process. In addition to the ACS, other response ARS elements B1, B2, B3 are also essential and found to be closely linked to the replication machinery. They are variable form of ACS with mutations at few positions and their mechanism can vary a lot across various organisms, which further complicates the computational prediction of replication region. Replication is obviously the most fundamental and essential process in the cell cycle of bacteria, and it is also one of major factor in exerting genome-wide mutational and selection pressure, shaping genomic polarity with asymmetrically biased nucleotide composition in leading and lagging strands [9]. The replication process initiate at origin of replication and that can be found with the help of Autonomous Replicating Sequence (ARS). The origin of replication is related to ARS sequence so genome sequence provides the foundation for future studies of organism, and is being exploited in

the search for new biological information. The parasite *Plasmodium falciparum* is responsible for hundreds of millions of cases of malaria, and kills more than one million African children annually [4]. Here we report an ARS analysis of the genome sequence of *Plasmodium falciparum* clone 3D7. The 23-megabase nuclear genome consists of 14 chromosomes, encodes about 5,300 genes. In the case of Yeast ARS play important role to mutations which constructed across the entire *Saccharomyces cerevisiae* chromosomal origin, ARSI [5]. Functional studies of these mutants disclose one essential element (A1), which includes a match to the ARS consensus sequence, and three additional elements (B1, B2, and B3), which collectively are also essential for origin function. ARS sequences are related to the origin of replication and each replication process includes ARS sequence. So for any biological replication ARS sequences are very important. To study life cycle of any organism we have to also know about their replication process, how they copy to each other and generate new same copy hence, computational study of ARS sequence is very important. Use of ARS based computation along with pattern search algorithms limit the search space so this approach is yielding results for analyze sequence data related to bacteria, archea and eukaryotic genomes. Replications in eukaryotes begin at particular sites known as origins of replication, or replicators. These replication origins occur throughout the genome, however the propensity of their occurrence depends on the type of organism. The prominent attribute of eukaryotic replication origins are best understood in the budding yeast *Saccharomyces cerevisiae*, where some ARS elements, confer origin activity. ARS elements are short DNA sequences of a few base pairs, recognize by their efficiency at initiating a replication event when cloned in a plasmid. Actually cellular origins of replication are poorly understood in most eukaryotes, as well as prokaryote. “Short chromosomal sequences have been cloned in the yeast *Saccharomyces cerevisiae* that enable plasmids to replicate along with the cellular chromosomes in the S phase of each cell cycle [10]”.

1.2 Objective

The major objective of our study is to make use of whole genome with special emphasis on sequence of few genomes including *Plasmodium falciparum* and computationally predict probable origin locations, study its organization and compare the same with ARS model organism.

1.3 Overview of thesis

chapter 2 provide some needed biological background of context. It deals with some introductory idea about plasmid, ARS sequences and some basic terminology about eukaryotes and prokaryotes.

Chapter 3 has two sections in which first section include about some basic description of computational tool which are implemented (PERL ,R, Linux, MATLAB) while second section give detail about methods used in this project.

Chapter 4 contain analysis and result in detail.

Chapter 5 provide brief summery and conclusion resulting from our study.
Some sample source code given in appendix in the end.

CHAPTER-2

BACKGROUND AND LITERATURE REVIEW

2.1 Plasmid and ARS.

The term plasmid was first introduced by the American molecular biologist Joshua Lederberg in 1952. Plasmids are directly related to ARS sequence and it is essential to study about its nature and functionality. A linear or circular double-stranded DNA that is capable of replicating independently of the chromosomal DNA.

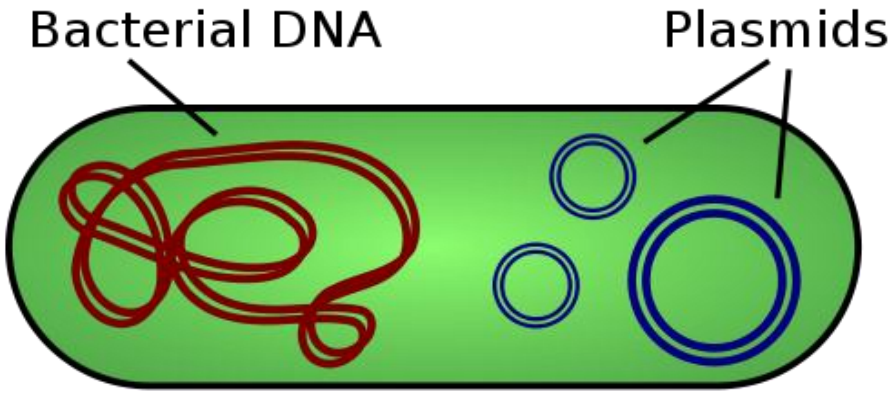


Figure 2.1.1: Illustration of a bacterium with plasmid enclosed showing genomic DNA and plasmids.
[Source: http://en.wikipedia.org/wiki/Plasmid]

Certain plasmids are able to insert themselves into the chromosomes particularly in regions where there is a common sequence of nucleotides are present. Hence, they are used in recombinant DNA technology and research as means of transferring genes between cells or used as cloning vectors. Plasmids are important in certain bacteria since plasmids code for proteins, especially enzymes, which offer resistance to antibiotics. Plasmids are very common in prokaryotes but they have also been found in a number of eukaryotes as well, e.g. *Saccharomyces cerevisiae*, which contain a 2-micrometre-ring of plasmid. Plasmid related to ARS sequence because Autonomously Replicating Sequence (ARS) elements are the genetic determinants of replication origin function in yeasts. They can be easily identified as the plasmids containing them appear in yeast cells at a higher frequency. The high number of plasmids generally decreases the growth rate, possibly allowing cells with few plasmids to dominate the culture.

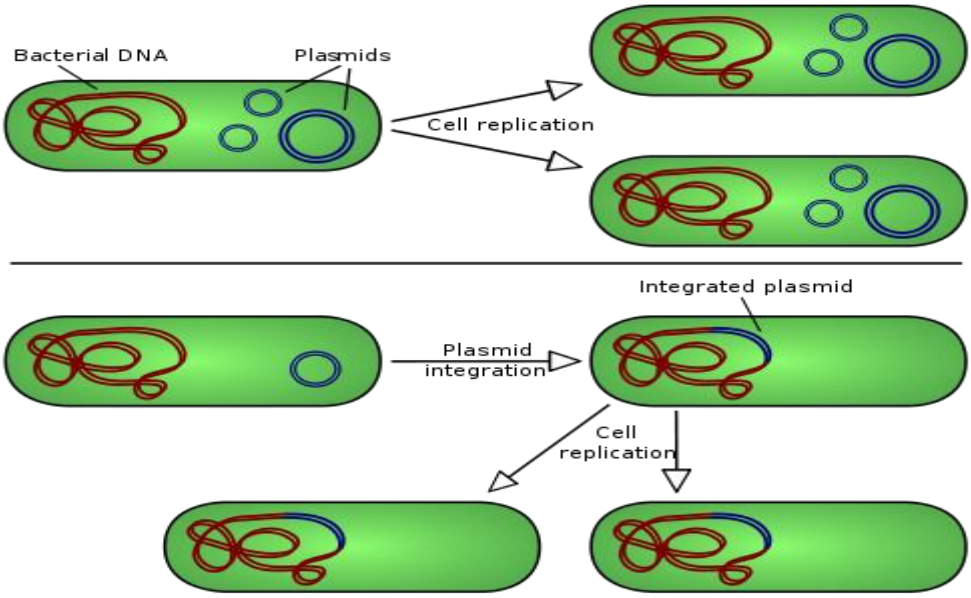


Figure 2.1.2: Two types of plasmid Non-integrating plasmids, Integrate plasmids [Source: http://en.wikipedia.org/wiki/Plasmid]

Plasmid vectors were classified into different classes based on their replication origin, selection marker and promoter information. The replication origins of plasmid vectors were classified as: prokaryotic replication origin, eukaryotic replication origin and viral replication origin[27].

2.2 Origin Of replication

The localized segment of genome where the replication process initiated is called origin of replication. Replication process is essential in living organisms such as prokaryotes and eukaryotes, or that of RNA or DNA in viruses, such as double-stranded RNA viruses. DNA replication may proceed either bidirectional or unidirectional. The specific structure of the origin of replication may vary somewhat from species to species, but all share some common salient property, such as high AT content. Origin of replication binds the pre-replication complex, a protein complex, which recognizes initiation site, unwinds, and begins to copy DNA. For smaller DNAs, including small viruses and bacteria plasmids single origin is sufficient. Larger DNAs have numerous origins, and DNA replication is initiated at all of them. If all replication process had to move from a single origin, it would take too long to replicate the entire DNA. Origin of replication also regulates the plasmid's characteristics: i.e. its ability to replicate in conjunction with another plasmid within the cell.

2.3 *Saccharomyces cerevisiae* and ARS

Saccharomyces cerevisiae is a species of yeast[11] and generally studied as a model organism. It is the most useful yeast, which is important to baking and brewing from ancient times. It is conceived that it was originally extracted from the skin of grapes yeast and also a constituent of the thin white film on the skins of some dark-colored fruits such as plums. It is one of the most widely analyzed eukaryotic model organisms in molecular biology, like Escherichia coli as the model bacterium. S. cerevisiae cells are either round or ovoid, 5–10 micrometers in diameter. It multiplies by budding mode. Many important proteins in human biology were first found through scientific observation by studying

their homolog's in yeast; these proteins include, signaling proteins, cell cycle proteins and protein-processing enzymes. There are two forms in which yeast cells can continue to live and grow: haploid and diploid. The haploid cells go through a simple life cycle of mitosis and growth, and under conditions of high stress. The diploid cells are the preferential 'form' of yeast also undergo a simple life cycle of mitosis and growth, but under conditions of high stress can undergo sporulation, go through meiosis and produce four haploid spores, which then can proceed on to mate. In *Saccharomyces cerevisiae* Autonomously replicating sequence-binding factor-1 (Abf1p) is an necessary sequence-specific transcription factor that takes part in multiple nuclear events including DNA replication, transcription activation, and gene silencing [12]. Many gene-specific probe have implicated Abf1p as a global transcriptional regulator involved in a diverse range of cellular functions.. Fragments of DNA with the properties expected for replication origins have been cloned from yeast chromosomal DNA. These sequences are found in the yeast genome at a frequency that is consistent with the spacing of origins estimated from electron microscopy and DNA fiber autoradiography, and they function to maintain plasmids autonomously in yeast cells.

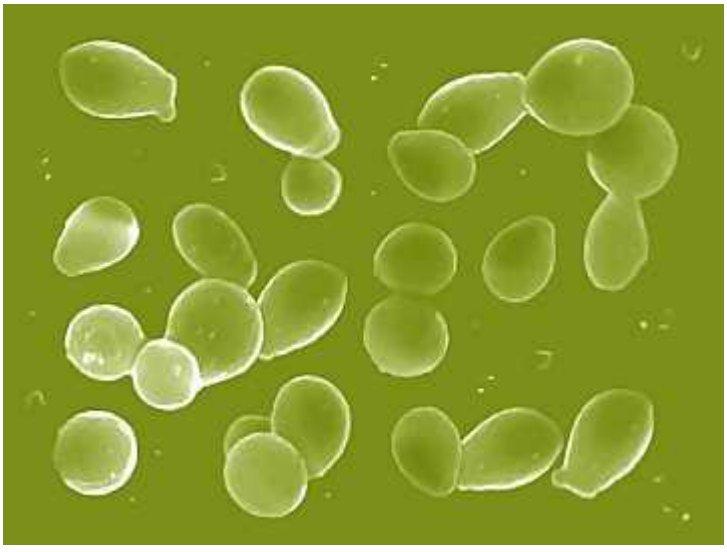


Figure 2.3.1: *Saccharomyces cerevisiae*

[Source:http://home.comcast.net/~pholowko/OnLineShows/Soil/MicroBio/BioSaccharomycesDescription.html]

2.4 Eukaryotes And Prokaryotes

Eukaryotes :-

Eukaryotes are an organisms whose cells contain complex structures confined within membranes. The presence of a nucleus imparts eukaryotes their name, which comes from the Greek (eu, "true") and (karyon, "nut" or "kernel"). Most eukaryotic cells have other membrane-bound organelles such as chloroplasts; mitochondria and the Golgi apparatus. Eukaryotes include organisms that all people are most familiar with e.g. plants, animals, fungi. They also include the majority of the organisms that palaeontologists deal with. Although they show incredible diversity in form, they share common characteristics of cellular organization, biochemistry and molecular biology. Here, the genetic material is packaged into chromosomes and is enveloped by a specialized membrane. The resulting structure is called the nucleus. Eukaryotic cells are much bigger - 10-100 μm - and exhibit a much more diverse and complex internal organization than prokaryotic cells Cell division in eukaryotes is unlike from that in organisms without a nucleus and more distinct. The basic eukaryotic cell contains plasma membrane , cytoplasm (semi fluid), glycocalyx, cytoskeleton.

Prokaryote :-

Prokaryotes include the domains of bacteria and archaea. Prokaryotes are single-celled organisms. They are the smallest and simplest organisms. They are abundant in the water, soil, air, and on most objects. The prokaryotes are a group of organisms that lack a cell nucleus (karyon), or any other membrane-bound organelles. The word prokaryote comes from the Greek (pro-) "before" + (karyon) "nut or kernel". Prokaryotes do not have a nucleus, mitochondria, or any other membrane-bound organelles. Prokaryotes belong to major taxonomic domains: the bacteria and the archaea. Archaea were recognized as a domain of life in 1990. These organisms were originally thought to live only in inhospitable conditions such as extremes of temperature, pH, and radiation but have since been found in all types of habitats. Prokaryotes are single-celled organisms that are the

earliest and most naive forms of life on earth. As organized in the Three Domain System, prokaryotes include bacteria and archaeans. Prokaryotes are able to live and thrive in various types of environments including extreme habitats such as hydrothermal vents, swamps, wetlands, hot springs, and the guts of animals. Prokaryotic cells are not as complex as eukaryotic cells. Following structures can be found in bacterial cells:
capsule, cell wall, cytoplasm Cell Membrane or Plasma membrane, Pili, Flagella Ribosomes Plasmids Nucleiod Region

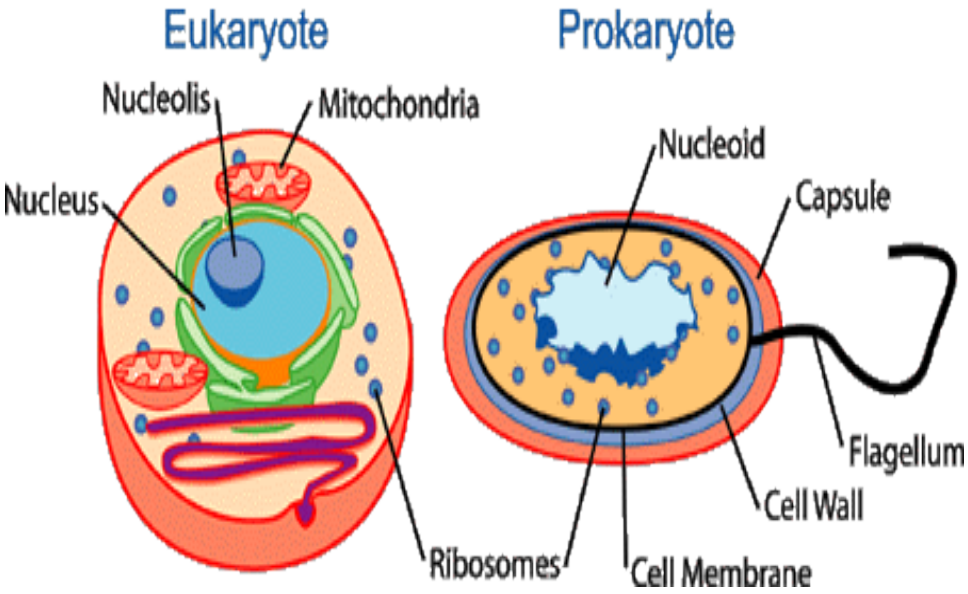


Figure 2.4.1: Basic structure of Eukaryotes and prokaryotes
[Source: http://www.ncbi.nlm.nih.gov/About/primer/genetics_cell.html]

2.5 *Plasmodium falciparum*

Plasmodium falciparum is a protozoan parasite, one of the species of Plasmodium that cause malaria in humans. It is transmitted by the female Anopheles mosquito. Malaria caused by this species is the most dangerous form of malaria with the highest rates of complications and mortality[30]. About 24 mega bases genome is reported and is organized into 14 chromosomes: just over 5,300 genes were described. *Plasmodium falciparum* is one of the four distinct species of the malaria parasite that affect humans and is responsible for 85 percent of the malaria cases. The parasite *Plasmodium falciparum* is responsible for hundreds of millions of cases of malaria, and kills more than one million African children annually[4].

2.6 Schizosaccharomyces pombe

Schizosaccharomyces pombe, also called "fission yeast", is a species of yeast. It is used as a model organism in molecular and cell biology[29]. It is a unicellular eukaryote, whose cells are rod formed. Cells normally measure 3 to 4 micrometers in diameter and 7 to 14 micrometers in length. Its genome, which is around 14.1 million base pairs These cells preserve their shape by growing exclusively through the cell tips and divide by medial fission to produce two daughter cells of equal sizes, which makes them a powerful tool in cell cycle research. Fission yeast was isolated in 1893 by Lindner from East African millet beer. The species name is derived from the Swahili word for beer (Pombe). It was first developed as an experimental model in the 1950s: by Urs Leupold for studying genetics and by Murdoch Mitchison for studying the cell cycle. The fission yeast researcher Paul Nurse successfully merged the the independent schools of fission yeast genetics and cell cycle research. Together wi th Lee Hartwell and Tim Hunt, Nurse won the 2001 Nobel Prize in Physiology or Medicine for their work on cell cycle regulation[29]. There are three chromosomes available in NCBI.

2.7 *Plasmodium berghei*

Plasmodium berghei is a unicellular parasite (protozoan) that infects mammals other than humans[28].The whole genome of *P. berghei* has been sequenced and it shows a high resemblance , both in structure and gene content, with the genome of the human malaria

Figure:2.5.1: [*Grammomys surdaster*, a natural host of *P. berghei*]
[Source http://www.lumc.nl/con]

parasite *Plasmodium falciparum*.P. berghei is found in the forests of Central Africa, where its natural cyclic hosts are the thicket rat (Grammomys surdaster) and the mosquito (Anopheles dureni). *P. berghei* is a part of a group of four *Plasmodium* species that infect murine rodents from Central Africa. These species are *P. vinckei, P. chabaudi, P. yoelii and P. berghei*. The first discovery of *P. berghei* was made by Vincke and Lips in 1948. *P. berghei* has been isolated from three different species of thicket rats. *P. berghei* infect hamsters, rats and mice. The life cycles and the different developmental stages of all malaria parasites are highly comparable. There are 14 chromosome available of *P. berghei*

CHAPTER-3

Materials and Methods

3.1 Some detail about computational tool which is used

3.1.1 PERL

3.1.2 R

3.1.3 MATLAB

3.1.4 Shell command and Linux command

3.2 Detail About methods which are Used:-

3.2.1 GC –Skew:-

3.2.2 Correlation C_G :-

3.2.3 Pattern Search with constraints:-

3.2.4 Genome wide exhaustive Pattern search:-

3.2.5 Distance distribution of ARS

3.1.1 PERL

In this project most of the coding is done in PERL programming language so a brief introduction is given required providing some detail about PERL. It is generally also used for string manipulation and string operation. For bioinformatics application there is BioPERL tool is available. PERL is a UNIX based language and was produced in 1987 by Larry Wall [31]. It was earlier created for UNIX based systems and it step by step evolved into a powerful tool for programming language originally for text manipulation regular expression. Regular expressions are a construct borrowed from automata theory. In computing, a regular expression provides a brief and flexible means to "match" strings of text, such as unique or specific characters, words, or patterns of characters. A regular expression, often called a pattern, is an expression that determines a set of strings. So far, we can seen easy use of scalar and list data is in PERL, but we have yet to explore the core of PERL's text processing construct regular expressions. Common shortened form for “regular expression” includes regex and regexp. We can use any text editor (notepad, vi) to write your PERL scripts or also we can use GUI application like eclipse or GEANY editor. A PERL file must be save with a .pl extension. The file can contain letters, numbers and symbols but must not contain a space. The other extensions you may come across are .ph - PERL .pm - PERL module header .pod - PERL documentation. The main features of PERL are , simple to learn, it is free, concise and easy to read ,fast, extensible. PERL has flexible data types, object oriented. PERL has a rich library of functions. They're the verbs of PERL, the commands that the interpreter runs. Bioinformatics, a rapidly evolving discipline, is the application of computational tools and techniques to the management and analysis of biological data. The term bioinformatics is relatively new, and as defined here, it encroaches on such terms as "computational biology" and others [18]. In biological research area there are huge amount of genome data and for their feature extract.

3.1.2 R

In this project R programming was used extensively. A brief introductory detail about R as given. R is a statistical computer program, made available under the General Public License (GPL). It is supplied with a license that allows you to use as the user requirement like Freely use , distribute it, as long as the receiver has the same rights and the source code is freely available. It exists for Unix and Linux platforms, Microsoft Windows 95 or later, for a variety of and for the Apple Macintosh (OS versions newer than 8.6).R provides an environment in which you can perform statistical analysis and produce graphics. For graphical purpose there are many packages available e.g. tcltk2, ggplot2. The R language is a project designed to create open source language totally free originally developed as the S language at AT&T Bell Labs. R is a highly functional language; package oriented which allow to use and call packages very easily and virtually most of things in R are done through functions. R provides two types for functions for graphics: high level functions, which produce an entire plot with a single call, and low-level functions, which are used to add additional information to existing graphs.

3.1.3 MATLAB

In this project for data manipulation and few step process MATLAB has been used so there is introduction of MATLAB is necessary and very concise detail is given. The name MATLAB stands for MATrix LABoratory. MATLAB was written originally to provide easy access to matrix software developed by the LINPACK (linear system package) and EISPACK (Eigen system package) projects. MATLAB is a high functioning language for technical computing. It incorporates computation, visualization, and programming environment. Furthermore, MATLAB is a modern programming language environment: it has advanced data structures, contains built-in editing and debugging tools, and supports object-oriented programming.

3.1.4 Linux and shell command

The Linux operating system, developed through the cooperation of many peoples around the world, and a product of the Internet and is a free operating system. In other words, all the source code is free. we are free to study it, redistribute it, and modify it. As a result, the code is available free of cost no charge for the software, source, documentation. The Linux operating system has many unique and powerful features. Like other operating systems, Linux is a control program for computers [19]. It supports many users. Linux provide Shell command. In my project I used shell command for some purpose. In PERL program shell command some were used.

3.2.1 GC Skew method

The GC skew method was the first computational method proposed for identification of origin of replication in genomes [2]. For a given sequence of nucleic acids, the GC skew measure is given by equation (1) where n_C and n_G are the number of occurrences of Cytosine (C) and Guanine (G).In this method, the origin of replication is said to be at the position where S undergoes an abrupt transition across S = 0.

Compositional asymmetries are spread throughout in DNA sequences. They are the result of the mutations arising from cells mechanisms such as replication and transcription. GC-skews methods were first used to study mitochondrial strand asymmetry [1]. Given two nucleotides G and C, with frequencies n_g and n_c, then their skew value S is defined as:-

$$S = \frac{n_c - n_g}{n_c + n_g}$$

eq. ……1

Procedure and methodology for GC skew method.

Count the total no of Cytosine (C) and Guanine (G) in few kilo bases. Use some windows size and slide this window towards starting to end of whole genome sequence. Applying above equation (1) and store position of each windows S value. This up to end of sequence. Plot S value verses window.

Genomic strand asymmetry phenomenon can be visualized using GC skew graphs. GC skew graph isolate the genome sequence into two segments: one with an excess of C over G corresponding to the lagging strand and the other with. an excess of G over C corresponding to the leading strand,

Pictorial view of GC skew method will be as follows.

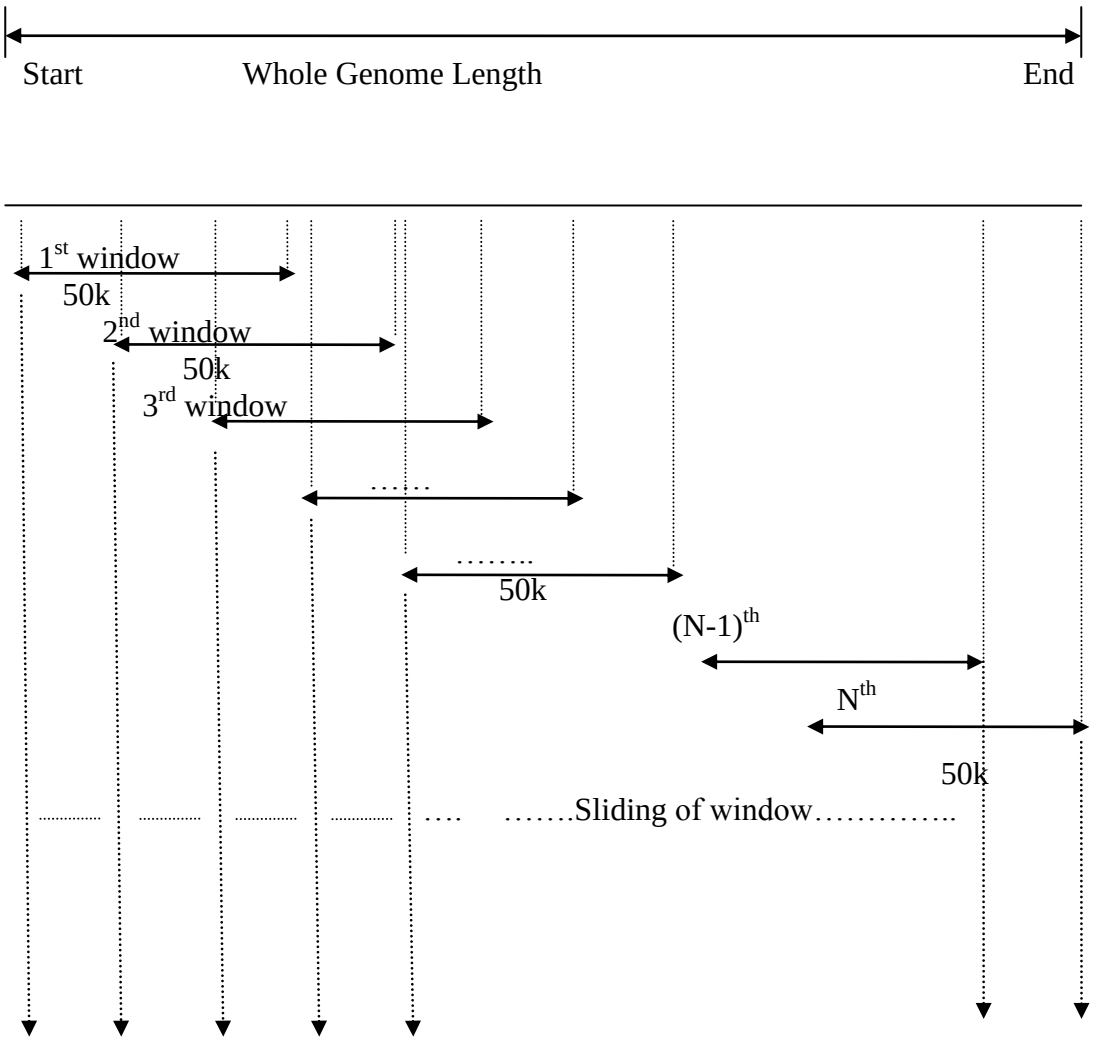


Figure 3.2.1.1: logical diagram of GC skew method

S = value stored in each iteration by applying above equation (1)

For each window total no of Cytosine (C) and Guanine (G) is counted and finding the value for S. This is done by sliding windows method. For each iteration next windows starting with new position

Window Stating position = Previous Window position + Slide Value

3.2.2 Correlation C_G method:-

It is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal which has been buried under noise, or identifying the missing from fundamental frequency in a signal implied by its harmonic frequencies. It is often used in signal processing for analyzing functions or series of values, such as time domain signals. To find correlation we multiply two functions (or the function by itself) at different values of the parameter and then take average over the domain of the function keeping the difference between those parameters: analyzing the correlation of a quantity will give information on how fast it changes and if it is somehow self-similar. The auto-correlation function, C(k), of a discrete sequence, $\{a_i : i = 1, 2, 3, 4 \ldots .,N\}$
with $a_i \in \{ + 1, -1\}$, The correlation function[6] can therefore be written as:-

$$C(k) = \frac{1}{N-k} \sum_{j=1}^{N-k} a_j a_{j+1} \qquad \ldots\ldots\ldots\ldots \text{Eq. (1),}$$

By Eq. (1), the correlation measure, C_G, can now be defined as the average of all correlation values as mentioned above in Eq. (1),

$$C_G = \frac{1}{N-1} \sum_{k=1}^{N-1} |C(k)| \qquad \text{..Eq.(2)}$$

where the subscript symbol “G” indicate to “genome sequence ”. The value for C_G ranges from zero to one [0,1] and is autonomous of the length of the genome sequence. Lower value of C_G corresponds to lower correlation strength embedded in that sequence and vice-versa. The value of C_G for a typical random sequence will be zero and a highly correlated sequence will approach unity. To use Eq. (2), we need to convert the nucleic acid sequence into a discrete sequence of bits. Since a DNA sequence is made up of four bases, we can generate a string of bits for the A base by assigning a value of +1 to every occurrence of A and -1 to all other positions (similarly for T, G, C). For example, a DNA sequence TCAGATGT gives rise to four different discrete sequences

{1,-1, -1, -1, -1, 1, -1,1} { -1,1,-1,-1,-1, -1, -1,-1}

{-1, -1, 1, -1, 1,-1,-1, -1} {-1, -1, -1, 1, -1, 1, 1,-1}

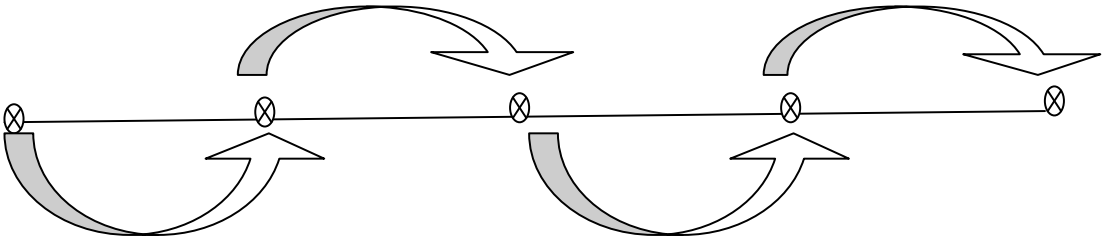
corresponding to the four bases A, T, G, C respectively. Thus, a given DNA sequence gives rise to four different bit strings and four different values of correlation strength (i.e. auto-correlation values) corresponding to each of the four bases, A, T, G, C. However, for the purpose of identifying the origin of replication, the correlation values of the Guanine residue give good results. In this method, it was found the origin of replication can be identified by an abrupt change in the value of correlation measure $C_{G.}$ The main distribution in that correlation is consider in genome while in GC nucleotide bases are identical as independent. Hence Correlation based models are more close to biological model.

Pictorial view of Correlation C_G method will be as follows:-

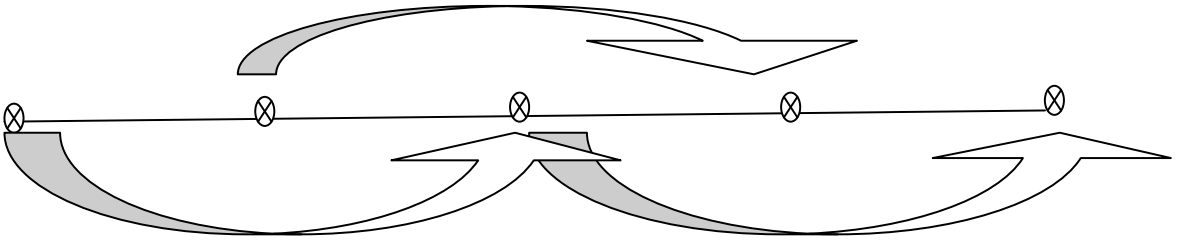
1 $$C(k) = \frac{1}{N-k}\sum_{j=1}^{N-k} a_j a_{j+k}$$

2 $$C_G = \frac{1}{N-1}\sum_{k=1}^{N-1} |C(k)|$$

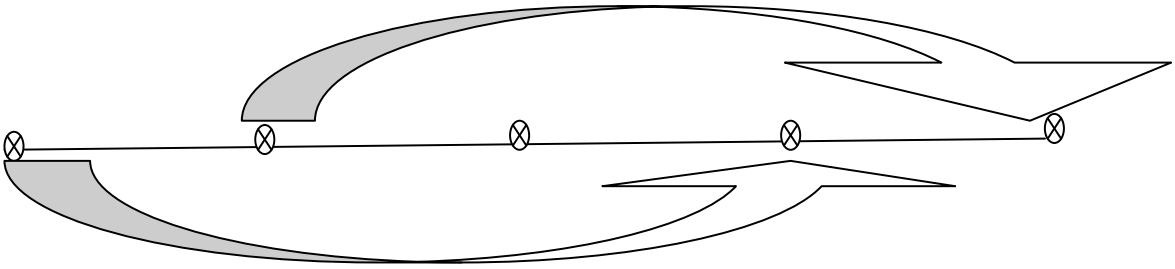
For N=5 and k=1 where N is data points, k is neighborhood position in figure shown bellow :-



For k=1; the pictorial representation of C(1) will be as above
There are five data points and each data points is multiply by another next value with **j=1**



For k=2; the pictorial representation of C(2) will be as above
There are five data set and each data set is multiply to another next value with j=1



For k=3; the pictorial representation of C(3) will be as above
There are five data set and each data set is multiply to another next value with j=1

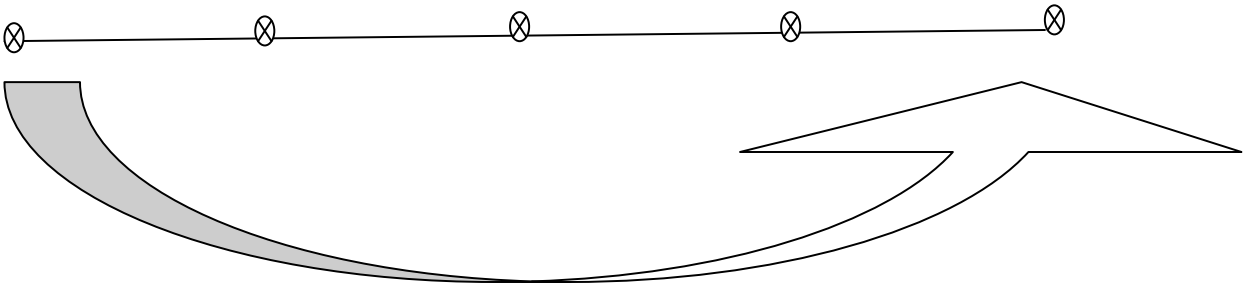


Figure 3.2.2.1: logical diagram of correlation method.

For k=4; the pictorial representation of C(4) will be as above
There are five data set and each data set is multiply to another next value with j=1

Now for getting each C(k) value with mod ,added together taking average of them will give correlation measure.

3.2.3 Pattern Search with constraints.:-

System and methods

In pattern search method , It is the Pattern Search algorithm which is able to analyze user submitted sequence collections for the presence of ARS patterns including potential secondary pattern elements also allowing mismatch below a user fixed threshold. The PatSearch program is written in PERL language and runs under the Unix/window operating system. It is essentially based on the pattern-matching program 'scan for matches' which is user input or embedded in program. The new version which implements the simulation procedure for assessing the statistical important of pattern hits are output .csv file or .txt file.

Implementation

The pattern Search program takes as input from fasta file format which include nucleotide bases [Downloaded from the server site Genbank, EMBL,NCBI and others] which is available in same directory . The users are allowed to choose whether they wish search sequence with some constraints. To search for nucleotide, whether they wish to search on both exact match and with some mismatch pattern of nucleotide sequence. Sequences, the maximum number of hits reported with exact match and mismatch for nucleotides {A,C,G,T }. The Pattern Search program locates all sequences position from the input sequences that are matched by a specified pattern[28]. The pattern description was inspired by 'regular expression' rules, although both the syntax and the semantics are different, especially for the inclusion of specific operators for finding regular expression. Here, It is clarify what we mean by a pattern and how the program locates the sequences matched position in output generated file.

The search pattern are mainly 4 types :-

Exact match ; one mismatch ; Two mismatch ; Three mismatch

Search sequence is " WTTTAYRTTTW"

Where :-

W = A or T

Y = C or T

R= A or G

Constraint imposed on above sequence and conditions are following.

a) Allow zero mismatch i.e. Exact match
b) Allow one mismatch
c) Allow two mismatch
d) Allow three mismatch

One mismatch means: - At one place of search string match with any string other than. the prescribed consensus pattern e.g. let string match to be ATTTACGTTTA and now all possible match will be as follows:-

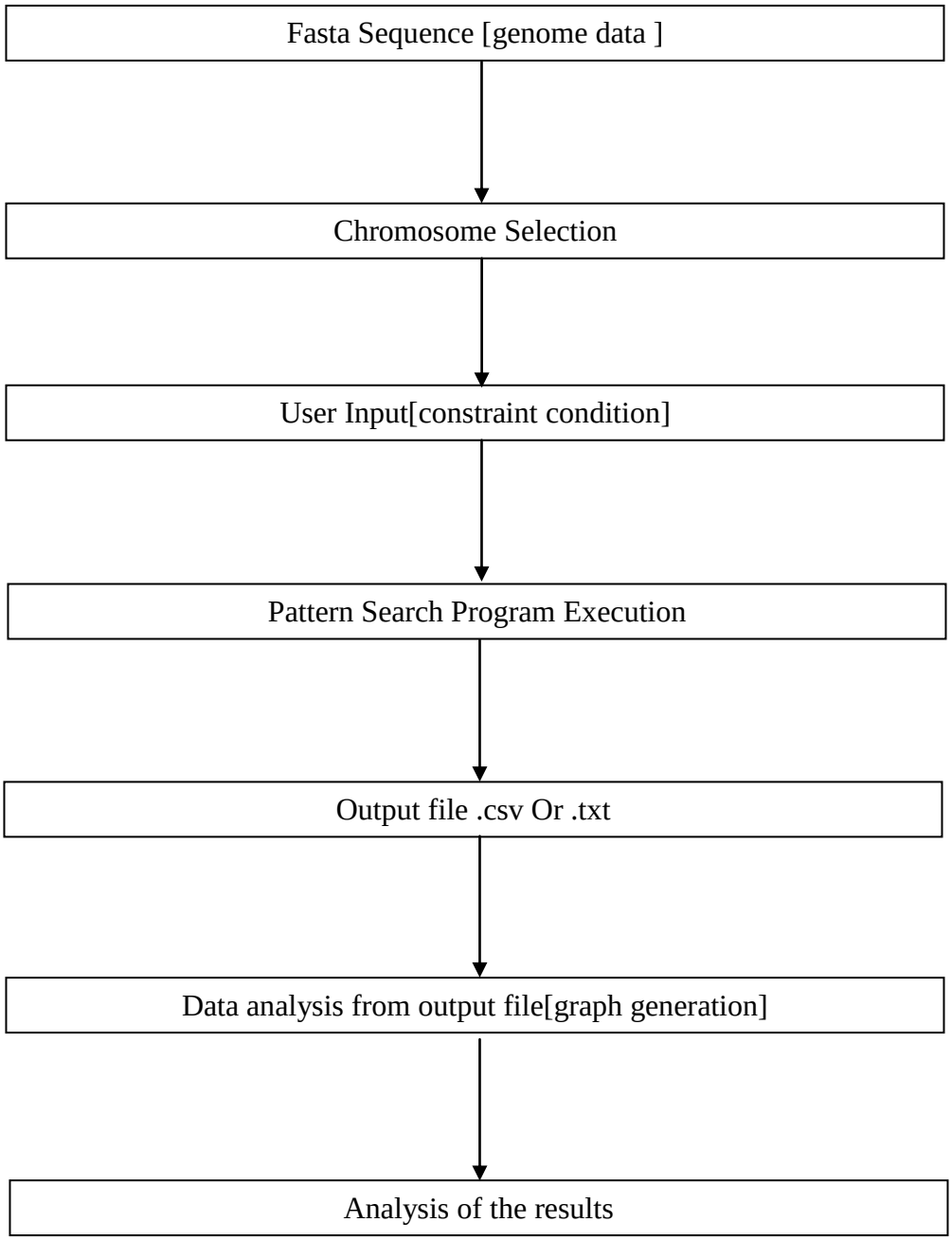
Table 3.2.3.1: Possible ARS pattern with one mismatch

1	*TTTACGTTTA	7	ATTTAC*TTTA
2	A*TTACGTTTA	8	ATTTACG*TTA
3	AT*TACGTTTA	9	ATTTACGT*TA
4	ATT*ACGTTTA	10	ATTTACGTT*A
5	ATTT*CGTTTA	11	ATTTACGTTT*
6	ATTTA*GTTTA		

Here "*" match with any string . For example in above table string *TTTACGTTTA match with

A TTTACGTTTA or TTTTACGTTTA or GTTTACGTTTA or CTTTACGTTTA

Pictorial view of Pattern Search method is as follows:-



3.2.4 **Genome wide exhaustive Pattern search :-**

For comparative study of ARS sequence of 11 length sequence, generation of 4^11(4194304) sequences and finding the sequences of 4^11 possibilities in each chromosome individually. Prepare top 4 frequency count table in comparison of ARS element of A1,B1,B2,B3

ALGORITHM:-

ARS element having length 11 and for sequence of 11 length string we can put each place 4 string. Therefore total no of string will be 4^11.

Logical Diagram for above concept is as bellow:-

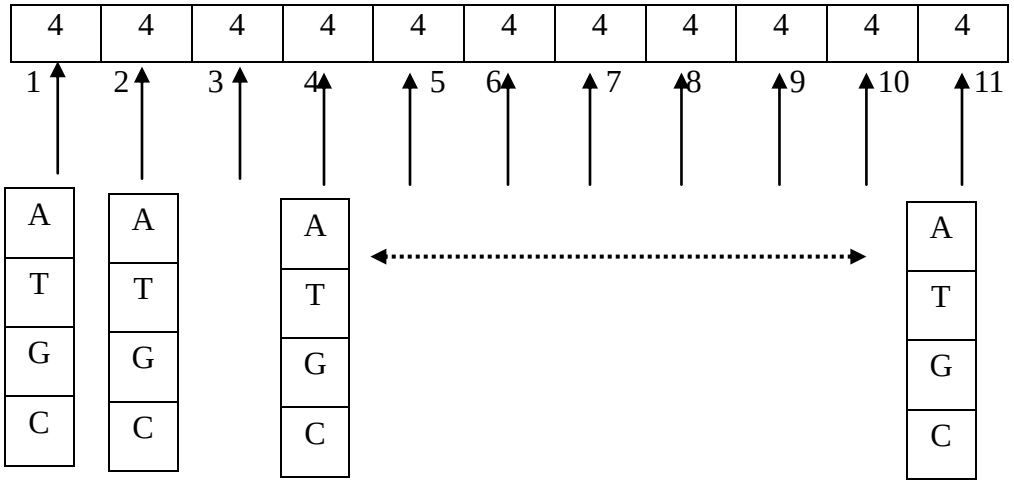


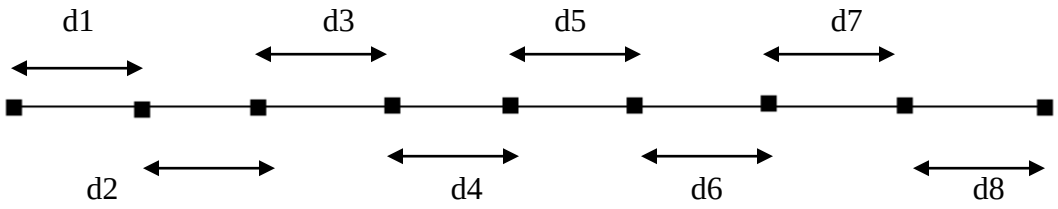
Figure 3.2.4.1: logical diagram of exhaustive Pattern search

In figure It is shown that at each position there are four possibilities of string and formation of 11 length sequence string each steps will be multiply together because according of Counting principle "If there are m ways to make a first selection and n ways to make a second selection, there are m × n ways to make the two selections. "

3.2.5 Distance distribution of ARS

In this section try to find the distance between two consecutive ARS sequences
This method may provide idea about ARS sequence distribution in genome sequences.

A pictorial view of this method mentioned bellow.



■ Indicate ARS Match Position

↔ Indicate Distance between two consecutive match sequences

d1= 2^{nd} match position – 1^{st} match position
d2= 3^{rd} match position - 2^{nd} match position

d^{n-1}= n^{th} match position – $(n-1)^{th}$ match position

CHAPTER-4

Results and Discussions

In this project the main objective is to explore the possibility of finding the ARS like sequences and make an attempt to identify origin of replication in few genomes ARS sequence and finding of origin of replication in few genomes. The Approach is through computational study with help of pattern search algorithms which reduce search space and find region of interest. Eexhaustive pattern search method is employed across the genome as described in materials and methods. In this study few genomes are selected for data analysis e.g. *Plasmodium falciparum, Saccharomyces cerevisiae, Plasmodium berghei* and *Schizosaccharomyces pombe*. For each genome, all chromosome have been examined. Result from pattern search program are visualized with suitable graphs. corresponding plot show some resulted which is discussed in proper section.

The following methods were implemented and used a genomic data:-

1.1 GC skew method

1.2 Correlation C_G

1.3 ARS Pattern Search with constraints

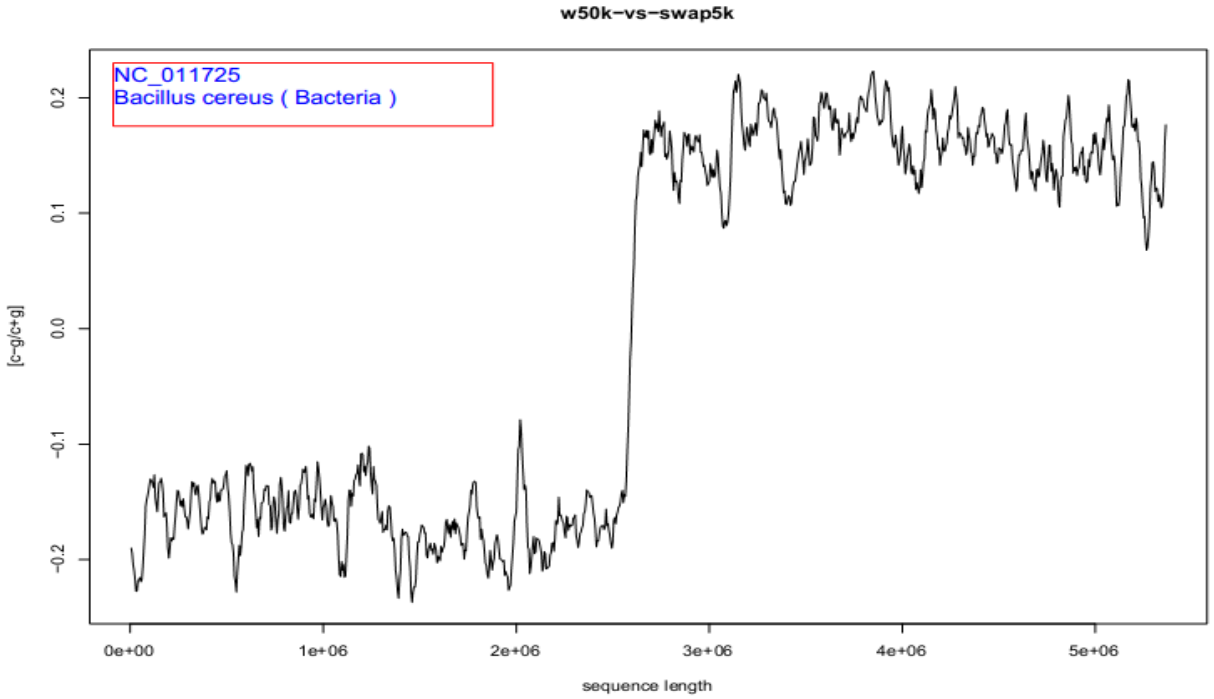
1.4 Genome wide exhaustive Pattern search

1.5 Distance distribution of ARS

4.1 GC skew method

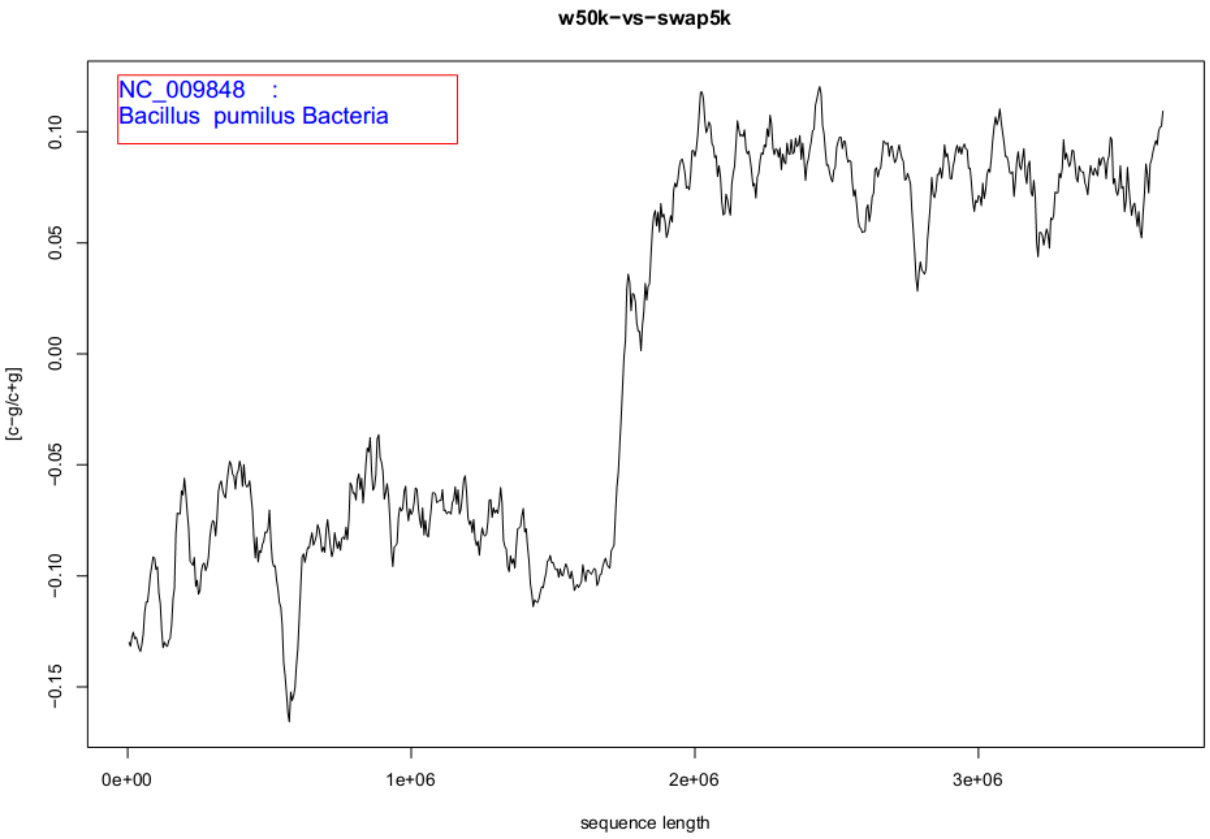
In GC skew method we computed the S values (compositional asymmetry) along the genome sequences. The purpose of GC skew method to identify the

the compositional asymmetry and look for the origin of replication. GC skew provide information about abrupt transition in 'S' values which points to possible sites. Our analysis shows GC skew method work fine for many bacterial genomes but fail in eukaryotic genomes. The reason may be due to the organization of origin sites.



origin of replication can be identified by an abrupt change in the value of S=0; where S = (nC-nG)/ (nC+nG)

Figure 4.1.1 : Origin of replication identification by GC skew method for Bacillus Cereus (Bacteria) having ACCESSION No: NC011725



origin of replication can be identified by an abrupt change in the value of S=0; where S = (nC-nG)/ (nC+nG)

Figure 4.1.2 : Origin of replication identification by GC skew method for Bacillus pumilus (Bacteria) having ACCESSION No: NC009848

Above plot shows that GC skew method is successful in identifying origin sites bacteria . Figure 4.3 and 4.4 shows the plot in the case of *P. falciparum* and *S. cerevisiae*.It is easily seen that the plot did not depict any particular pattern rather look like random behavior. Species wise [*Plasmodium falciparum* ,*Saccharomyces cerevisiae* ,*Plasmodium berghei* Schizosaccharomyces pombe]plot for all chromosomes are provided in appendix.

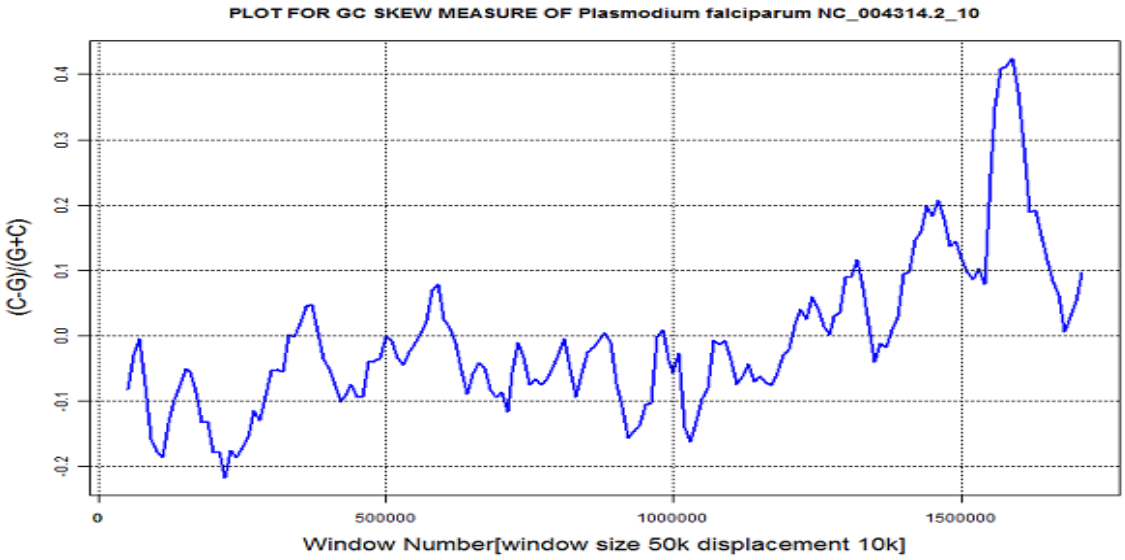


Figure 4.1.3 : Origin of replication identification by GC skew method for *P. falciparum* having ACCESSION No: NC004314.2

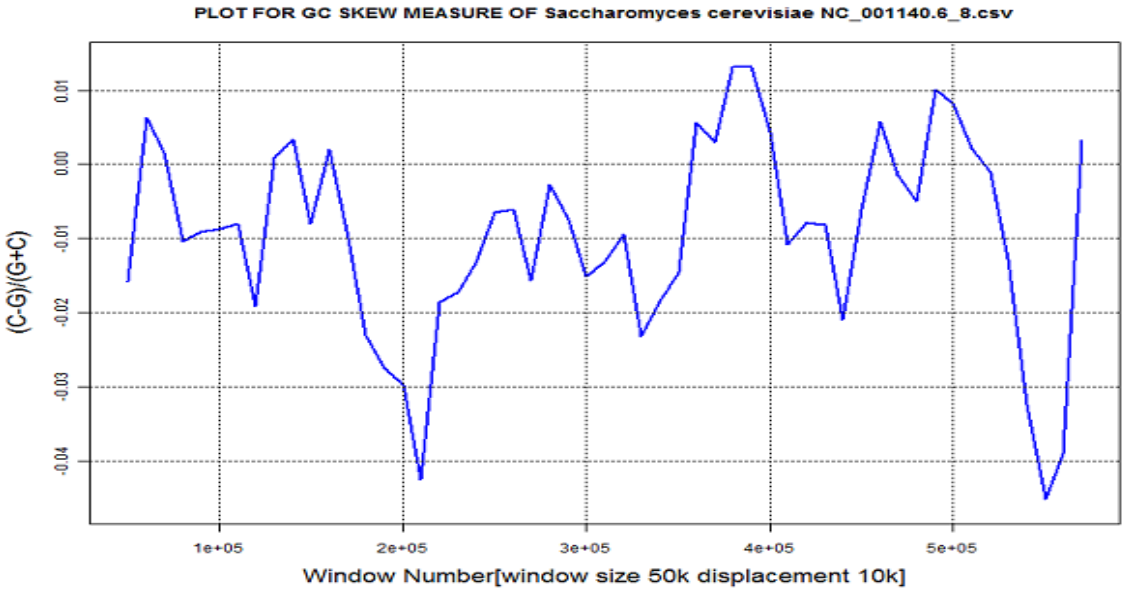


Figure 4.1.4: Origin of replication identification by GC Skew method for *Saccharomyces cerevisiae* having ACCESSION No : NC001140.6

4.2 Correlation C_G method

In Correlation C_G method we compute the correlations within the sequences.The value ranges from 0 to 1. The value of C_G for noisy or random sequence will be near to zero and more correlated sequence will be near about 1. In the correlation plot the origin of replication is seen by noting the abrupt change in the plotted values

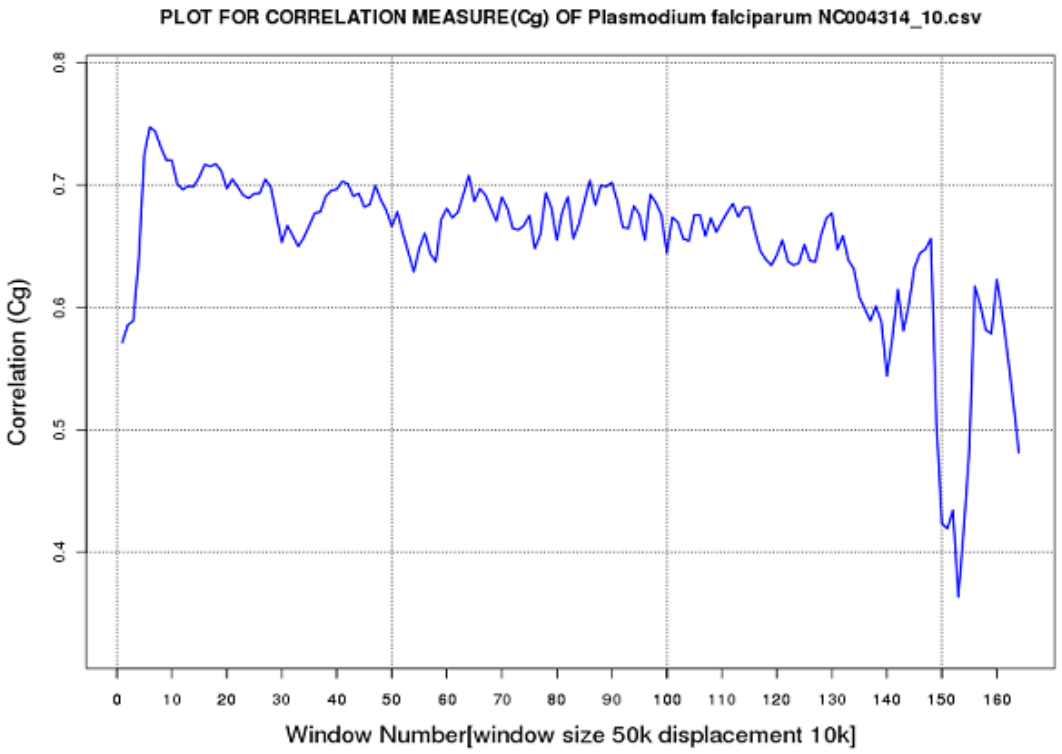


Figure 4.2.1: Origin of replication identification by Correlation C_G method for *P. falciparum* having ACCESSION No : NC004314[10^{th} chromosome]

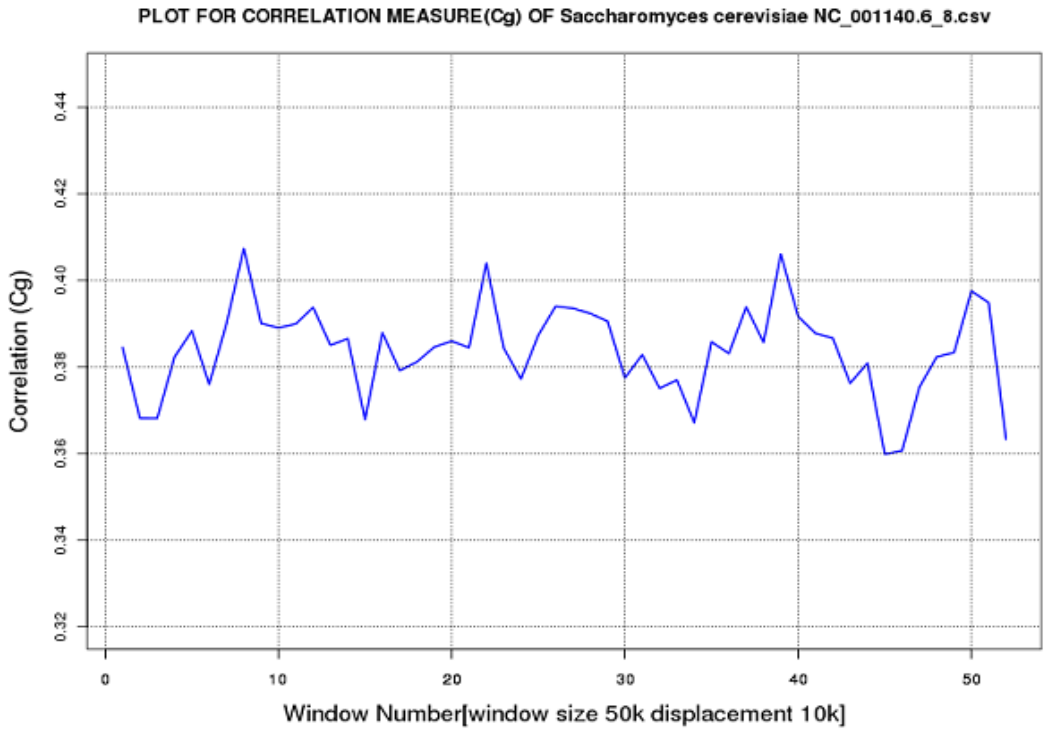


Figure 4.2.2: Origin of replication identification by Correlation C_G method for *Saccharomyces cerevisiae* having ACCESSION No : NC001140.6

In Figure 4.2.1 between windows no 150 to 160 the abrupt change is found so it indicates that it is partially identify origin of replication. Species wise [*Plasmodium falciparum* ,*Saccharomyces cerevisiae* ,*Plasmodium berghei* Schizosaccharomyces pombe] all chromosome detail for Correlation C_G method see the Figure appendix.

Here combined results for GC Skew and Correlation Measure has been shown in tabular form for species *Plasmodium falciparum ,Saccharomyces cerevisiae ,Plasmodium berghei ,Schizosaccharomyces pombe*.

Table describes the behavior of plots as non random and noisy.

Where,

1. Non Random : It state that graph has more minima and fluctuations have more valley points(minima)
2. Noisy : It depicts that graph has more or less minima with constant fluctuation

******* In table NR used for Non Random

Table 4.2.1

Prediction of Origin of Replication Finding *Plasmodium falciparum* 3D7 [Eukaryota]				
ACCESSION No	Method		Size (kb)	Remark
	GC -Skew $S = \frac{n_c - n_g}{n_c + n_g}$	Correlation C_G $C(k) = \frac{1}{N-k}\sum_{j=1}^{N-k} a_j a_{j+1}$		
NC_004325.1	NR	NR	628.10	
NC_000910.2	NR	NR	924.90	
NC_000521.3	NR	NR	1035.24	
NC_004318.1	NR	NR	1175.89	
NC_004326.1	Noisy	NR	1312.06	
NC_004327.2	Noisy	Noisy	1385.00	
NC_004328.2	Noisy	Noisy*	1466.52	In Correlation C_G some region found
NC_004329.2	Noisy	Noisy	1386.29	
NC_004330.1	Noisy	Noisy	1505.59	
NC_004314.2	Noisy	Noisy*	1648.10	In Correlation C_G some region found
NC_004315.2	Noisy	Noisy	1990.56	
NC_004316.3	Noisy	Noisy	2218.24	
NC_004331.2	Noisy	Noisy	2827.74	
NC_004317.2	Noisy	Noisy	3214.72	

Here table 4.2 give complete description of GC skew method and correlation measure method combined result analysis. It provides details about ***Plasmodium falciparum*** [all 14 chromosome] analysis. It also indicate that genome length and plot.

Table 4.2.2

Prediction of Origin of Replication Finding *Saccharomyces cerevisiae* [Eukaryota]				
ACCESSION No	Method		Size (kb)	Remark
	GC -Skew $S = \frac{n_c - n_g}{n_c + n_g}$	Correlation C_G $C(k) = \frac{1}{N-k}\sum_{j=1}^{N-k} a_j a_{j+1}$		
NC_001133.9	NR	NR	224.82	
NC_001134.8	Noisy	Noisy	794.13	
NC_001135.5	NR	NR	309.20	
NC_001136.10	Noisy	Noisy	1496.03	
NC_001137.3	Noisy	NR	563.35	
NC_001138.5	NR	NR	263.83	
NC_001139.9	Noisy	Noisy	1065.37	
NC_001140.6	NR*	NR	549.46	* In GC skew method some region found
NC_001141.9	NR	NR	429.57	
NC_001142.9	Noisy	Noisy	728.27	
NC_001143.9	Noisy	Noisy	651.19	
NC_001144.5	Noisy	Noisy	1052.91	
NC_001145.3	Noisy	Noisy	902.76	
NC_001146.8	NR	NR	765.95	
NC_001147.6	Noisy	Noisy	1065.71	
NC_001148.4	Noisy	Noisy	925.85	

Here table 4.2 give complete description of GC skew method and correlation measure method combined result analysis. It provide detail about *Saccharomyces cerevisiae* [all 16 chromosome] analysis. It also indicate that genome length and nature of plot.

Table 4.2.3

Prediction of Origin of Replication Finding *Plasmodium berghei* ; [Eukaryota]				
Chromosome No	Method		Size (kb)	Remark
	GC -Skew $S = \frac{n_c - n_g}{n_c + n_g}$	Correlation C_G $C(k) = \frac{1}{N-k}\sum_{j=1}^{N-k} a_j a_{j+1}$		
Chromosome 1	NR	NR	463.92	
Chromosome 2	NR	NR	621.44	
Chromosome 3	NR	NR	573.31	
Chromosome 4	NR	Noisy	707.25	
Chromosome 5	NR	Noisy	895.84	
Chromosome 6	NR	NR	896.89	
Chromosome 7	NR	NR	791.36	
Chromosome 8	Noisy	Noisy	1322.33	
Chromosome 9	Noisy	Noisy	1593.85	
chromosome 10	Noisy	Noisy	1542.58	
Chromosome 11	Noisy	Noisy	1678.78	
Chromosome12	Noisy	Noisy	1723.03	
Chromosome 13	Noisy	Noisy	2434.11	
Chromosome 14	Noisy	Noisy	2390.79	

Here table 4.3 give complete description of GC skew method and correlation measure method combined result analysis. It state about *Plasmodium berghei* [all 14 chromosome] analysis. It also indicate that genome length and nature of plot analysis

Table 4.2.4

Origin of Replication Finding *Schizosaccharomyces pombe* ;[Fungi]				
ACCESSION No	Method		Size (kb)	Remark
	GC -Skew $S = \frac{n_c - n_g}{n_c + n_g}$	Correlation C_G $C(k) = \frac{1}{N-k}\sum_{j=1}^{N-k} a_j a_{j+1}$		
NC_003424.3	Noisy*	Noisy	5527	*In GC skew method some region found
NC_003423.3	Noisy	Noisy	4497	
NC_003421.2	Noisy	Noisy*	2430	* Correlation C_G method some region found

Here table 4.4 gives complete description of GC skew method and correlation measure method result analysis. It state about *Schizosaccharomyces pombe* [all 3 chromosome] analysis. It also indicate that genome length and nature of plot.

4.3 ARS Pattern Search with constraints

ARS are important for replication process in eukaryote as well as in prokaryote.
In presence of high access of ARS in genome, is more significant for fast replication.
Here counting of ARS element for *Plasmodium falciparum, Saccharomyces cerevisiae, Plasmodium berghei, Schizosaccharomyces pombe* is done for all chromosome which are available. The present results give a comprehensive view of ARS element in genome.
Result also shows the distribution and ARS density in chromosome.
Search pattern are for types:-

- Exact ARS match pattern
- One ARS mismatch pattern
- Two ARS mismatch pattern
- Three ARS mismatch pattern

Exact match pattern give idea about more probability of origin of replication.
One mismatch ARS pattern indicate about likely origin of replication will be found. same as two or three mismatch pattern follow above trend. I.e. three mismatch ARS patterns are dubious about of origin of replication.

The detail about result is following in form of tables & graphs.
ARS sequence analyses in form of table are as follows:-

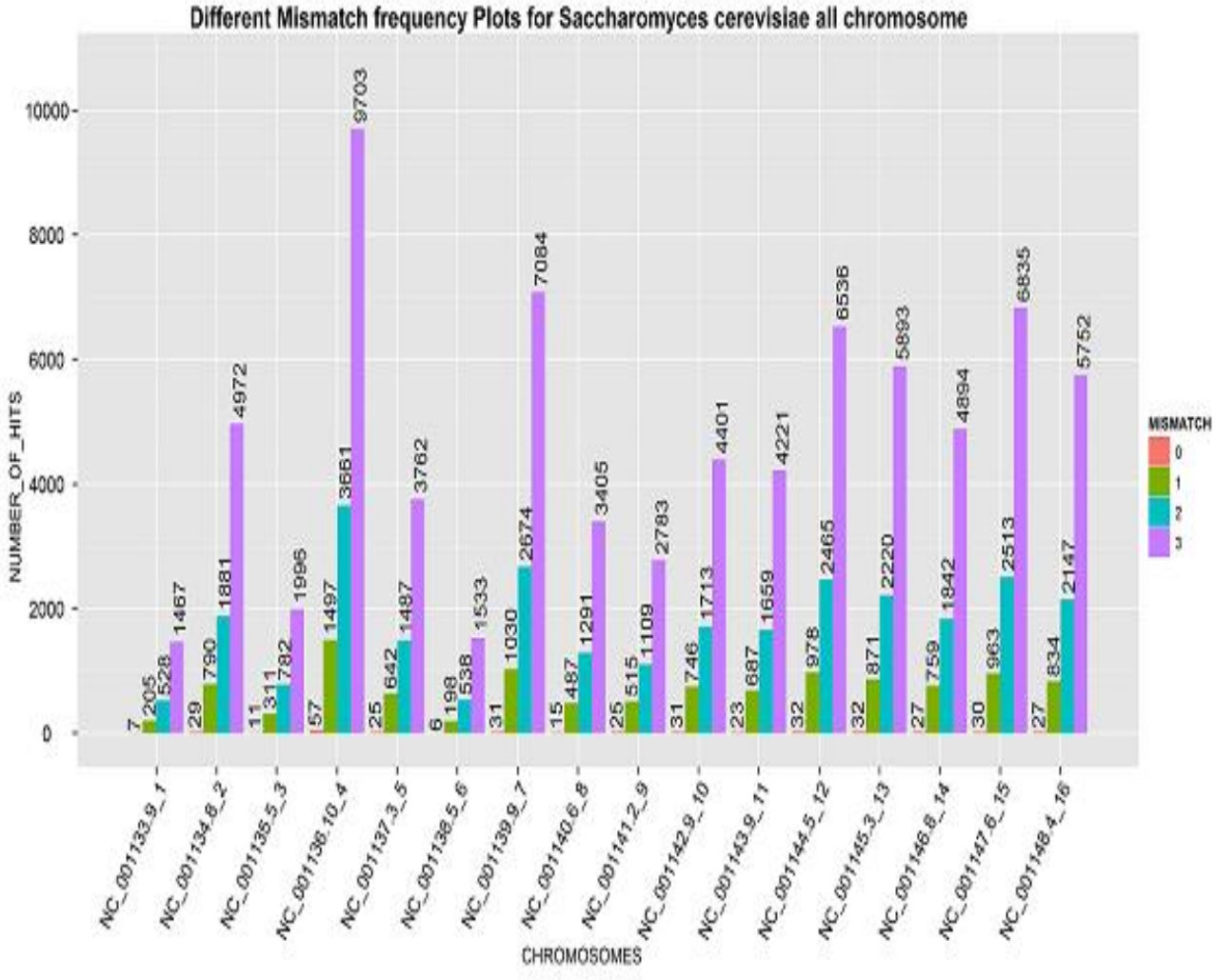


Figure 4.3.1: Plot of *Saccharomyces cerevisiae* ARS sequence match

With exact match (Zero mismatch), one mismatch and two mismatch and three mismatch. Plot gives complete idea about total no of ARS sequence match in all 14 chromosomes of Saccharomyces *cerevisiae*.

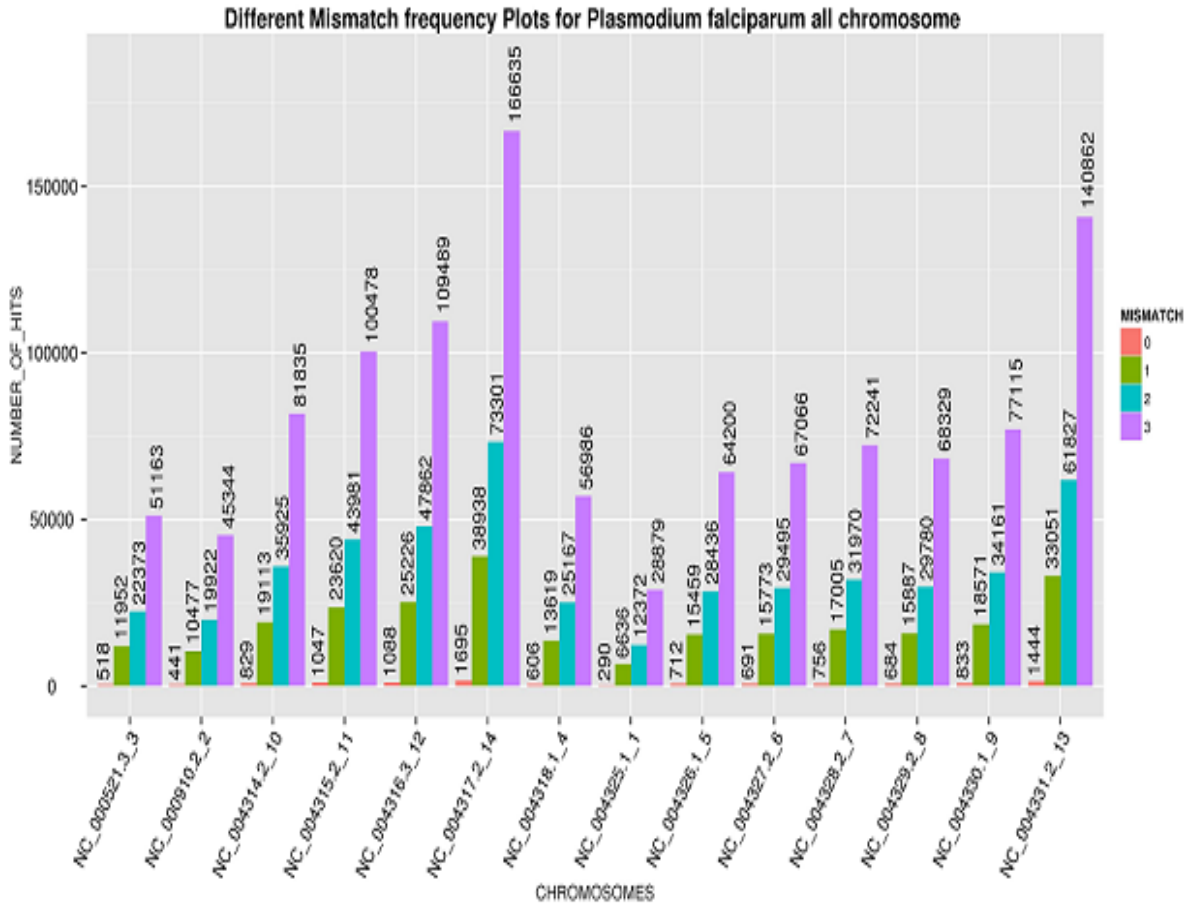


Figure 4.3.2: Plot of *Plasmodium falciparum* ARS sequence match with exact match (Zero mismatch), one mismatch and two mismatch and three mismatch. Plot gives complete idea about total no of ARS sequence match in all 14 chromosomes of *Plasmodium falciparum.*

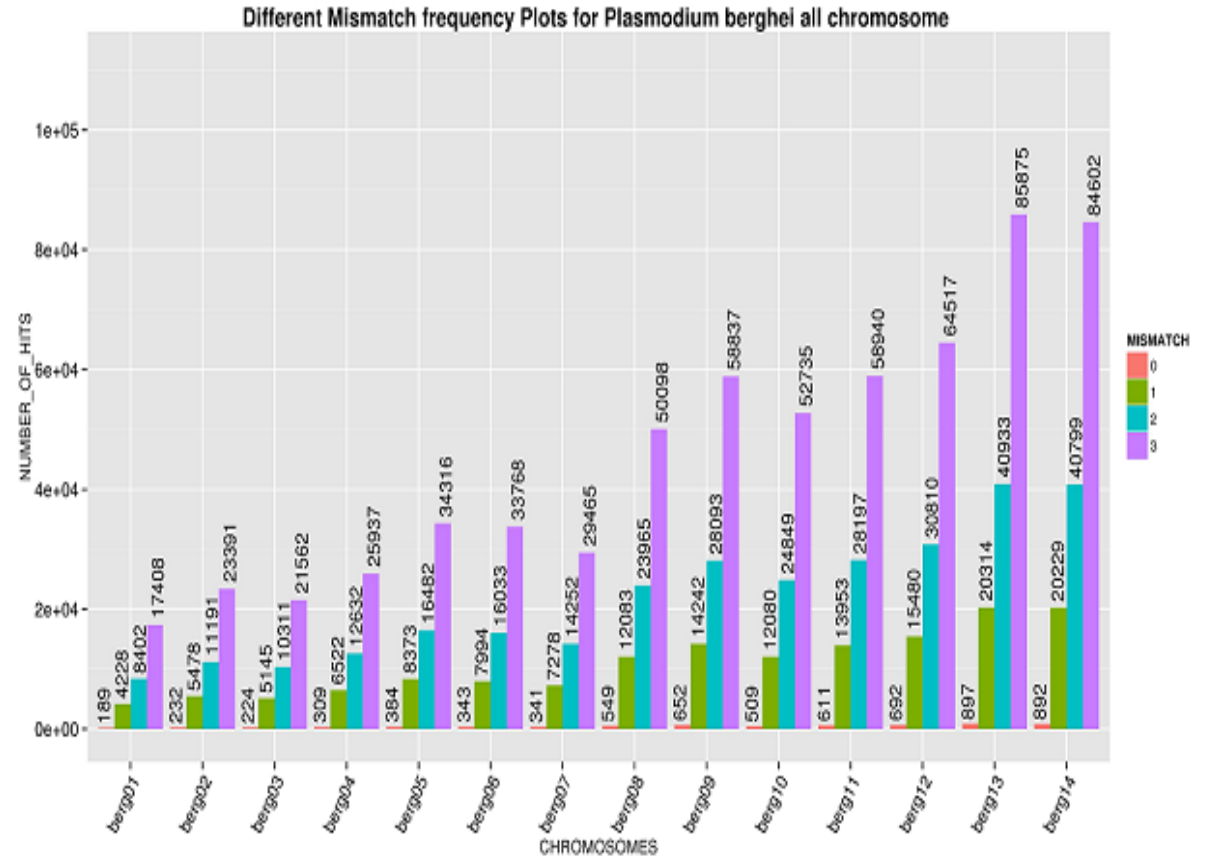


Figure 4.3.3: Plot of *Plasmodium berghei* ARS sequence match with exact match (Zero mismatch), one mismatch and two mismatch and three mismatch. Plot gives complete idea about total no of ARS sequence match in all 14 chromosomes of *Plasmodium berghei.*

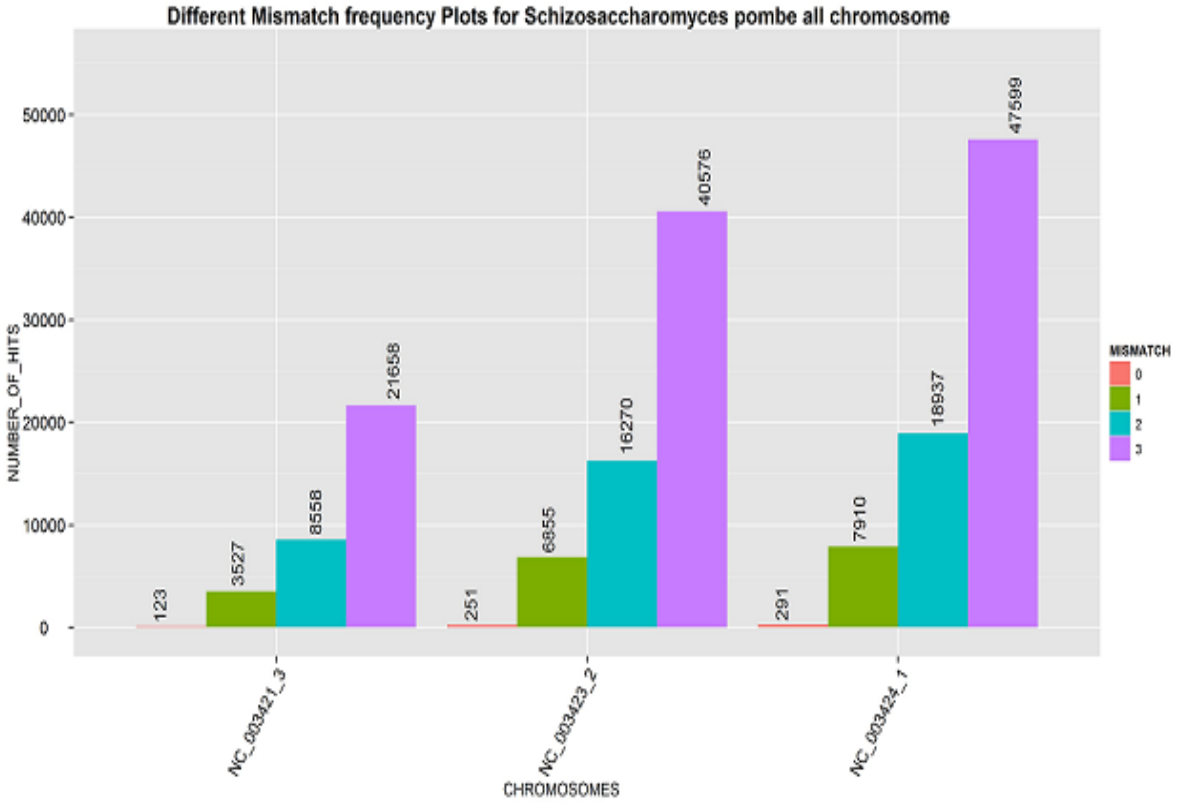


Figure 4.3.4: Plot of *Plasmodium berghei* ARS sequence match with exact match (Zero mismatch), one mismatch and two mismatch and three mismatch. Plot gives complete idea about total no of ARS sequence match in all 14 chromosomes of *Plasmodium berghei.*

Table 4.3.1

ARS analysis: - ORGANISM *Saccharomyces cerevisiae* Eukaryota; Fungi

$\frac{0[A1]}{SIZE}$ (density)	ACCESSION no	SIZE(kb)	Total No Of ARS pattern Hits			
			0[A1]	1[B1]	2[B2]	3[B3]
0.031136	NC_001133.9	224.82	7	205	528	1467
0.036518	NC_001134.8	794.13	29	790	1881	4972
0.035576	NC_001135.5	309.20	11	311	782	1996
0.038101	NC_001136.10	1496.03	57	1497	3661	9703
0.044377	NC_001137.3	563.35	25	642	1487	3762
0.022742	NC_001138.5	263.83	6	198	538	1533
0.029098	NC_001139.9	1065.37	31	746	1713	4401
0.0273	NC_001140.6	549.46	15	487	1291	3405
0.058198	NC_001141.9	429.57	25	515	1109	2783
0.042567	NC_001142.9	728.27	31	746	1713	4401
0.03532	NC_001143.9	651.19	23	687	1359	4221
0.030392	NC_001144.5	1052.91	32	978	2465	6536
0.035447	NC_001145.3	902.76	32	871	2220	5893
0.03525	NC_001146.8	765.95	27	759	1842	4894
0.02815	NC_001147.6	1065.71	30	963	2513	6835
0.029162	NC_001148.4	925.85	27	834	2147	5752

*1kb = 1024 bases [nucleotides]

Table 4.3.2

ARS analysis: - ORGANISM *Plasmodium falciparum* 3D7 Eukaryota

$\frac{0[A1]}{SIZE}$ (density)	ACCESSION No.	SIZE(kb)*	Total No Of ARS pattern Hits			
			0[A1]	1[B1]	2[B2]	3[B3]
0.46171	NC_004325.1	628.10	290	6636	12372	28879
0.476808	NC_000910.2	924.90	441	10477	19922	81835
0.500367	NC_000521.3	1035.24	518	11952	22373	51163
0.515354	NC_004318.1	1175.89	606	13619	25167	56986
0.542658	NC_004326.1	1312.06	712	15459	28436	64200
0.498917	NC_004327.2	1385.00	691	15773	29495	67066
0.515506	NC_004328.2	1466.52	756	17005	31970	72241
0.493403	NC_004329.2	1386.29	684	15887	29780	68329
0.553271	NC_004330.1	1505.59	833	18571	34161	77115
0.503003	NC_004314.2	1648.10	829	19113	35925	81835
0.525983	NC_004315.2	1990.56	1047	23630	43981	100478
0.490479	NC_004316.3	2218.24	1088	25226	47862	109489
0.510655	NC_004331.2	2827.74	1444	33051	61827	140862
0.527262	NC_004317.2	3214.72	1695	38938	73301	166635

*1kb = 1024 bases [nucleotides]

Table 4.3.3

ARS analysis: - ORGANISM *Plasmodium berghei Eukaryota; Fungi;*

$\frac{0[A1]}{SIZE}$ (density)	SIZE(kb)	Total No Of ARS pattern Hits			
		0[A1]	1[B1]	2[B2]	3[B3]
0.407398	463.92	189	4228	8402	17408
0.373326	621.44	232	5478	11191	23391
0.390714	573.31	224	5145	10311	21562
0.436903	707.25	309	6522	12632	25937
0.428648	895.84	384	8873	16482	34316
0.382433	896.89	343	7994	13033	33768
0.430904	791.36	341	7278	14252	29465
0.415176	1322.33	549	12083	23965	50098
0.409072	1593.85	652	14242	28093	58837
0.329967	1542.58	509	12080	24849	52735
0.363955	1678.78	611	13953	28197	58940
0.401618	1723.03	692	15480	30810	64517
0.368513	2434.11	897	20314	40933	85875
0.373098	2390.79	892	20229	40799	84602

*1kb = 1024 bases [nucleotides]

Table 4.3.4

ARS Detail About: - ORGANISM Schizosaccharomyces *pombe Eukaryota; Fungi;*

$\frac{0[A1]}{SIZE}$ (density)	ACCESSION no	SIZE(kb)	Total No Of ARS pattern Hits			
			0[A1]	1[B1]	2[B2]	3[B3]
0.46171	NC_003424.3	5527	291	7910	18937	47599
0.476808	NC_003423.3	4497	251	6855	16270	40576
0.500367	NC_003421.2	2430	123	3527	8558	21658

*1kb = 1024 bases [nucleotides]

From the above tables data shows about complete detail of the occurrence of ARS sequences in *Saccharomyces cerevisiae*, *Plasmodium falciparum, Plasmodium berghei and* Schizosaccharomyces *pombe.*. Tables show for each chromosome how many ARS match has been found [0 to 3 mismatch].

From table analysis it has been found that for *Saccharomyces cerevisiae and* Schizosaccharomyces *pombe* have more ARS sequences in comparison of *Plasmodium falciparum* and *Plasmodium berghei* in reference of per kb ARS i.e. ARS density.

$$\text{ARS density} = \frac{Total\ number\ of\ Exact\ Match\ ARS}{\text{chromosome length(kb)}}$$

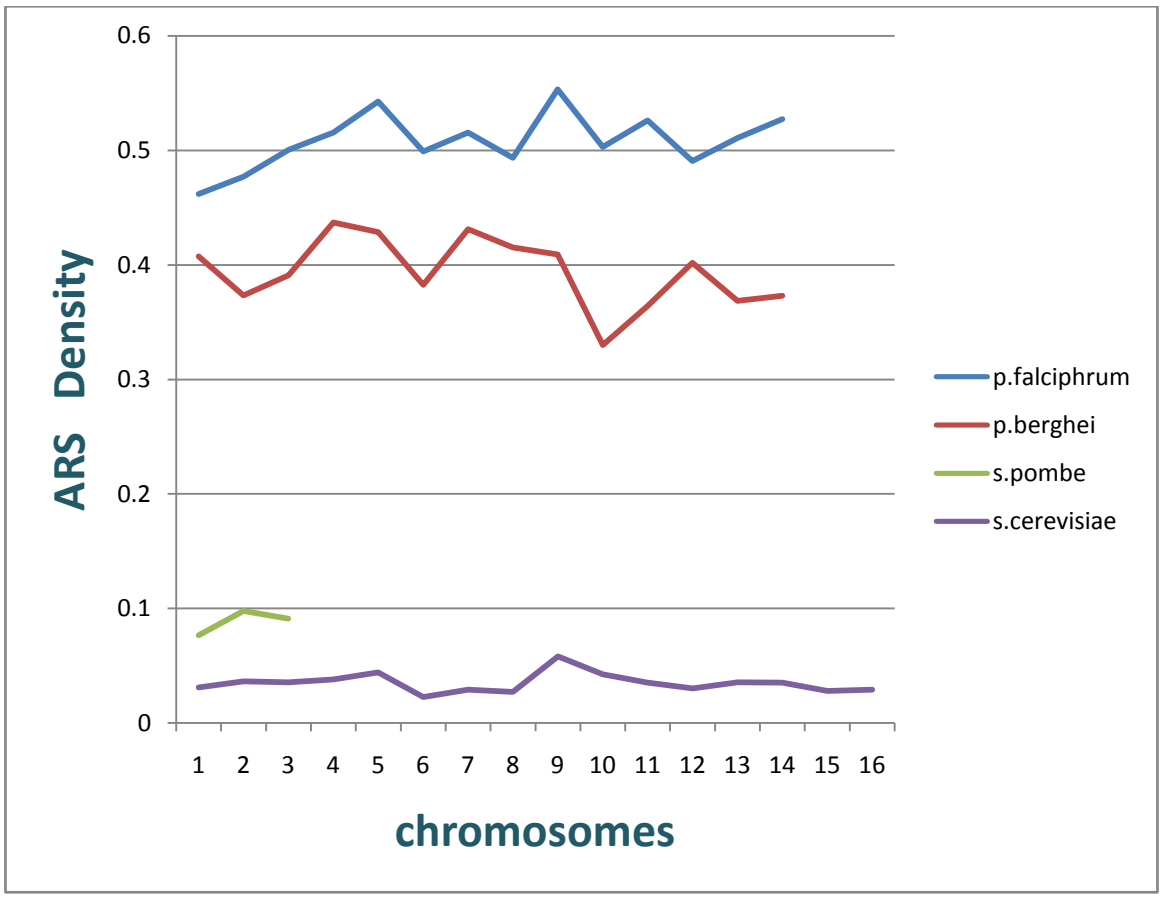


Figure 4.3.1 : Figure shows ARS density i.e. per kb total count of ARS in *Plasmodium falciparum, Saccharomyces cerevisiae, Plasmodium berghei, Schizosaccharomyces pombe* is done for all chromosomes.

Figure shows that there is few no of ARS in *Saccharomyces cerevisiae and Schizosaccharomyces pombe* but in case of *Plasmodium* it is high.

4.4 Genome wide exhaustive Pattern search

For comparative study of ARS sequence to all possible 11 length sequence. Analyzed all pattern which are all possible to construct with A,T,G, and C [ATGC] i.e. it's include 4^11 search sequences. For each pattern we find corresponding frequency in the genome..Graphical plot is shown bellow. All plot shows that they have some high no of pattern and almost its total no is four. Some plot is shown bellow.

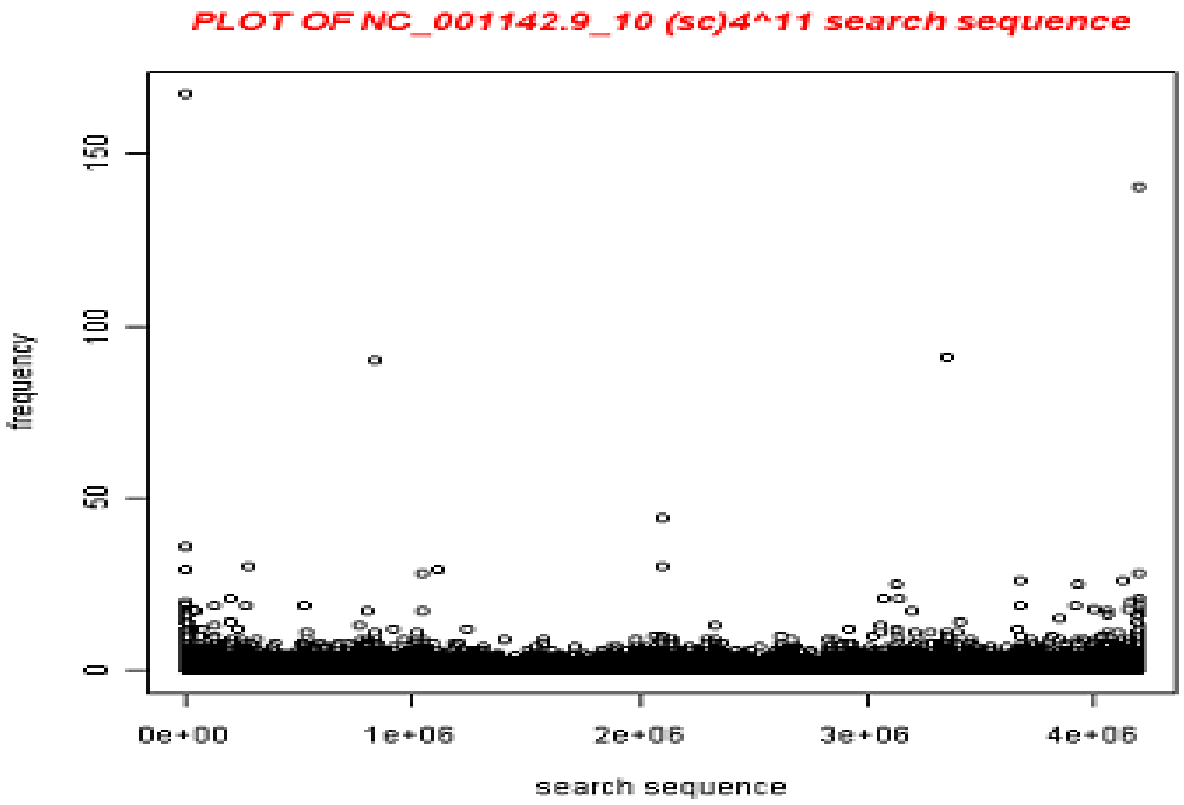


Figure 4.4.1: It plot of 4^11 search pattern in *Saccharomyces cerevisiae* 10th chromosome.

Genome wide exhaustive Pattern search result has been tabulated which is mentioned bellow. The analyses are shown in table .It provides top 4 hits of search sequence in genome and corresponding sequence pattern.

TABLE 4.4.1: 4^11 exhaustive Pattern Search for *Saccharomyces cerevisiae*

Accession No	Top 1st	Top 2nd	Top 3rd	Top 4th
NC_001133.9	106	85	21	18
	TTTTTTTTTTT	AAAAAAAAAAA	ATATATATATA	TATATATATAT
NC_001134.8	263	209	101	96
	AAAAAAAAAAA	TTTTTTTTTTT	ATATATATATA	TATATATATAT
NC_001135.5	111	83	78	63
	AAAAAAAAAAA	TATATATATAT	ATATATATATA	TTTTTTTTTTT
NC_001136.10	372	277	222	222
	TTTTTTTTTTT	AAAAAAAAAAA	ATATATATATA	ATATATATATA
NC_001137.3	207	116	71	68
	AAAAAAAAAAA	TTTTTTTTTTT	ATATATATATA	TATATATATAT
NC_001138.5	68	61	48	46
	TTTTTTTTTTT	AAAAAAAAAAA	ATATATATATA	TATATATATAT
NC_001139.9	312	291	161	158
	AAAAAAAAAAA	TTTTTTTTTTT	ATATATATATA	TATATATATAT
NC_001140.6	216	135	95	92
	AAAAAAAAAAA	TTTTTTTTTTT	ATATATATATA	TATATATATAT
NC_001141.2	113	111	54	50
	AAAAAAAAAAA	TTTTTTTTTTT	TATATATATAT	ATATATATATA
NC_001142.9	167	140	91	90
	AAAAAAAAAAA	TTTTTTTTTTT	TATATATATAT	ATATATATATA
NC_001143.9	152	121	107	101
	TTTTTTTTTTT	AAAAAAAAAAA	TATATATATAT	ATATATATATA
NC_001144.5	279	249	131	120
	AAAAAAAAAAA	TTTTTTTTTTT	ATATATATATA	TATATATATAT
NC_001145.3	260	156	118	117
	TTTTTTTTTTT	AAAAAAAAAAA	TATATATATAT	ATATATATATA
NC_001146.8	241	172	92	91
	TTTTTTTTTTT	AAAAAAAAAAA	TATATATATAT	ATATATATATA
NC_001147.6	307	211	120	116
	TTTTTTTTTTT	AAAAAAAAAAA	ATATATATATA	TATATATATAT
NC_001148.4	230	166	95	87
	AAAAAAAAAAA	TTTTTTTTTTT	ATATATATATA	TATATATATAT

TABLE 4.4.2: 4^11 exhaustive Pattern Search for *Plasmodium falciparum.*

Accession No	Top 1st	Top 2nd	Top 3rd	Top 4th
NC_004325.1	6810	6801	4465	4328
	ATATATATATA	TATATATATAT	TTTTTTTTTTT	AAAAAAAAAAA
NC_000910.2	10544	10503	6188	5905
	ATATATATATA	TATATATATAT	AAAAAAAAAAA	TTTTTTTTTTT
NC_000521.3	11656	11632	7068	6782
	TATATATATAT	ATATATATATA	TTTTTTTTTTT	AAAAAAAAAAA
NC_004318.1	12449	12435	7364	7305
	TATATATATAT	ATATATATATA	TTTTTTTTTTT	AAAAAAAAAAA
NC_004326.1	16132	16092	9684	8788
	TATATATATAT	ATATATATATA	AAAAAAAAAAA	TTTTTTTTTTT
NC_004327.2	15685	15672	9047	8811
	ATATATATATA	TATATATATAT	TTTTTTTTTTT	AAAAAAAAAAA
NC_004328.2	14710	14698	9443	9209
	ATATATATATA	TATATATATAT	TTTTTTTTTTT	AAAAAAAAAAA
NC_004329.2	16224	16205	9795	9724
	TATATATATAT	ATATATATATA	AAAAAAAAAAA	TTTTTTTTTTT
NC_004330.1	19655	19613	11485	10627
	ATATATATATA	TATATATATAT	AAAAAAAAAAA	TTTTTTTTTTT
NC_004314.2	19908	19855	11578	11003
	ATATATATATA	TATATATATAT	AAAAAAAAAAA	TTTTTTTTTTT
NC_004315.2	24712	24670	14011	13814
	TATATATATAT	ATATATATATA	TTTTTTTTTTT	AAAAAAAAAAA
NC_004316.3	26817	26751	16348	14818
	ATATATATATA	TATATATATAT	AAAAAAAAAAA	TTTTTTTTTTT
NC_004331.2	32960	32948	19226	18591
	ATATATATATA	TATATATATAT	TTTTTTTTTTT	AAAAAAAAAAA
NC_004317.2	38733	38668	22139	21845
	TATATATATAT	ATATATATATA	TTTTTTTTTTT	AAAAAAAAAAA

4.5 Distance distribution of ARS

Here we measure how ARS sequences have been distributed. In this section
It is calculated that what distance between consecutive ARS sequences is.
Our objective is to find to find how ARS is distributed over whole chromosomes.

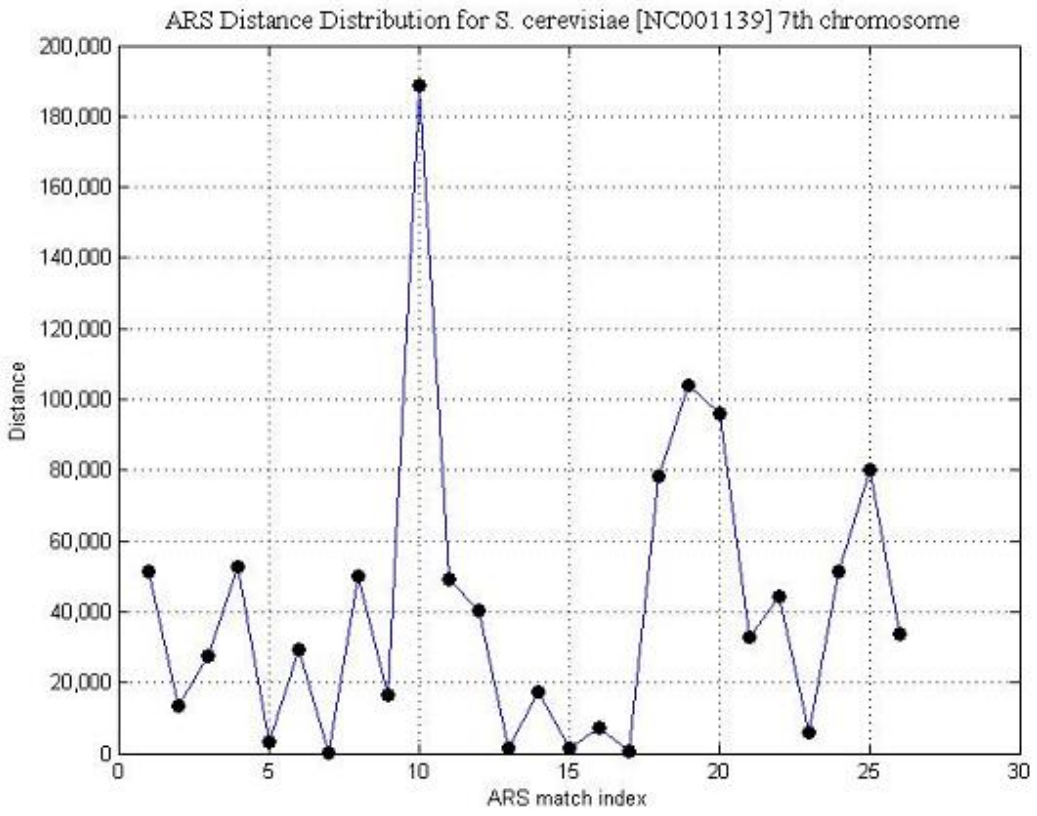


Figure 4.5.1: Plot of ARS consecutive match distance distribution for *S.cerevisiae* of 7^{th} chromosome.

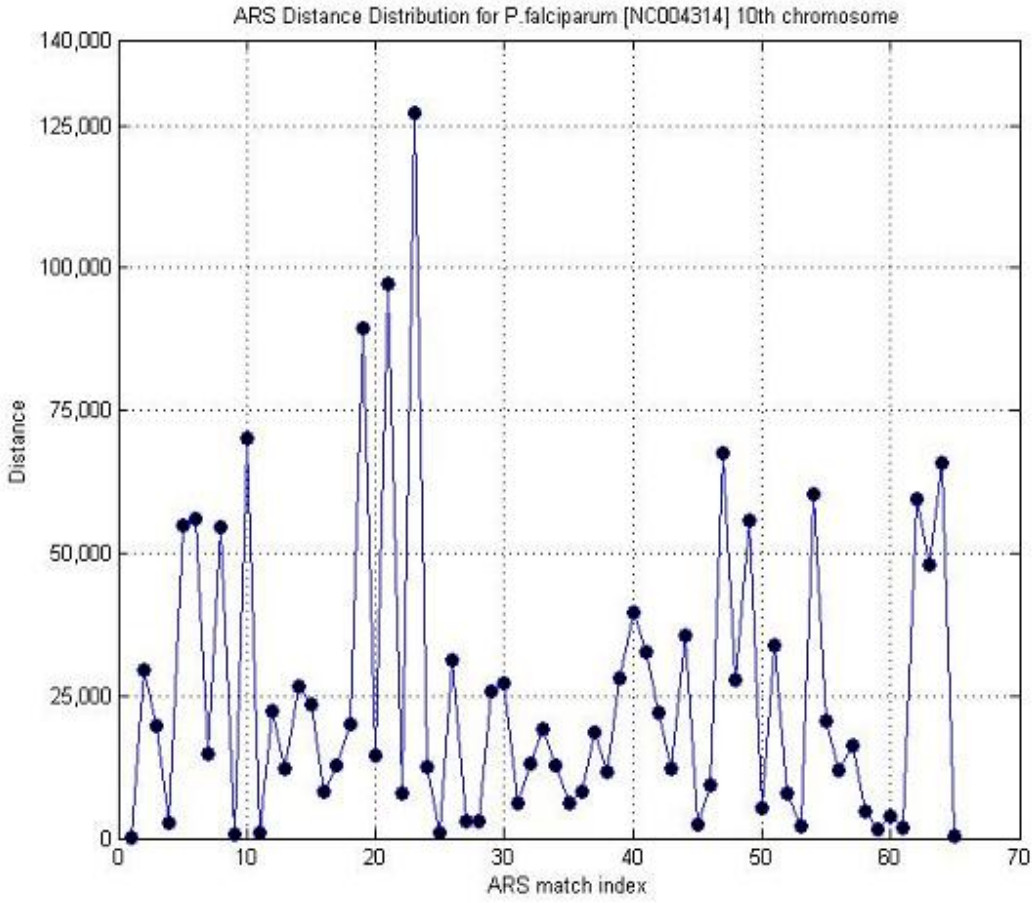


Figure 4.5.2: Plot of ARS consecutive match distance distribution for *P. falciparum* of 10^{th} chromosome.

Figure 4.5.1 and 4.5.2 state that most of ARS sequence is just distance of ~ [50kb] distance. Very few number of ARS which are consecutively found more than [100kb]

CHAPTER-5

Conclusion and Future Work:-

In this chapter, we summarize the results of our study about origin of replication in few genomes. The conclusions are as follows: -

The suitability of GC Skew and correlation method is explored for few non bacterial genomes i.e. *Plasmodium falciparum, Saccharomyces cerevisiae, Plasmodium berghei* Schizosaccharomyces *pombe* .

Result shows, above mentioned two methods could not clearly identify origin of replication. In few cases (for some chromosome) correlation method is better than GC skew. The reason could be that , origin sites are many and new innovative strategy has to be adopted. ARS like sequences were searched in whole genome of p.falciparum and p. bergh and the result were coupled with the model organism *Saccharomyces cerevisiae.* As expected no of exact match less than. Density of ARS element in *Saccharomyces cerevisiae* is less than the *Plasmodium falciparum.* Consecutive ARS pattern are found within the region of 50k.Few outliers are also these. Whole genome is searched for all contain of 11-mer nucleotides. Top from 11-mer pattern are identified. All top hits contain either A or T nucleotide (higher A+T).

Future work: - The organization of predicted origin sites has to be investigated through the neighborhood region (context based) analysis such as energy landscape chromatin to get more insights.

References :-

1. N.T. Perna, T.D. Kocher,[1995], Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes, *J. Mol.* E vol. **41(1995)** 353e358
2. Mrazek, J., Karlin, S., [1998]. Strand compositional asymmetry in bacterial and large viral genomes. *Proceedings of the National Academic Sciences of the United States* **95**, 3720–3725.
3. Kazuharu Arakawa and Masaru Tomita, [2007], The GC Skew Index: A Measure of Genomic Compositional Asymmetry and the Degree of Replicational Selection. *Evolutionary Bioinformatics*: **3** 159–168
4. Malcolm J. Gardner[1] Neil Hall[2] Eula Fung[3] Owen White[1], Matthew Berriman[2], Richard W. Hyman[3], Jane M. Carlton[1], Arnab Pain[2],Karen E. Nelson[1], Sharen Bowman[2]*, Ian T. Paulsen[1], Keith James[2], Jonathan A. Eisen[1], Kim Rutherford[2], Steven L. Salzberg[1], Alister Craig[4],Sue Kyes[5], Man-Suen Chan[5], Vishvanath Nene[1], Shamira J. Shallom[1], Bernard Suh[1], Jeremy Peterson[1], Sam Angiuoli[1], Mihaela Pertea[1],Jonathan Allen[1], Jeremy Selengut[1], Daniel Haft[1], Michael W. Mather[6], Akhil B. Vaidya[6], David M. A. Martin[7], Alan H. Fairlamb[7],Martin J. Fraunholz[8], David S. Roos[8], Stuart A. Ralph[9], Geoffrey I. McFadden[9], Leda M. Cummings[1], G. Mani Subramanian[10], Chris Mungall [11],J. Craig Venter[12], Daniel J. Carucci[13], Stephen L. Hoffman[13]*, Chris Newbold[5], Ronald W. Davis[3], Claire M. Fraser[1]& Bart Barrell[2].,Genome sequence of the human malaria parasite Plasmodium falciparum,[2002] *NATURE* | VOL **419** |
5. York Marahrens and bruce Stillman, [Feb. 14, 1992], A Yeast Chromosomal Origin of DNA Replication Defined by Multiple Functional Elements,. *Science* Vol. **255**, No. 5046 pp. 817-823.
6. Annangarachari Krishnamachari, Kushal Shah, [2012] ; Nucleotide correlation based measure for identifying origin of replication in genomic sequences; *BioSystems* **107** 52– 55

7. William J.Bagaria, Sonia M.F.Garcia ; Richard P. Fahey Rapid , [2004];Binary Gage Function to Extract a Pulsed Signal Buried in Noise. *EURASIP*:**13**, 1985–1992
8. McFadden,G.I.,Roos,D.S.,[1999].Apicomplexan plastids as drug targets.Trends in Microbiology , *Elsevier Science* **7**, 328–333.
9. Jean R Lobry, Noboru Sueoka [2002], Asymmetric directional mutation pressures in bacteria *Genome Biology*, 3**(10):**research0058.1–0058.14
10. R. T. Simpson, (1990), A transcriptionally active tRNA gene interferes with nucleosome positioning in vivo. *Nature* 343, 387
11. http://en.wikipedia.org/wiki/Saccharomyces_cerevisiae
12. Tsuyoshi Miyake, Justin Reese, Christian M. Loch, David T. Auble, and Rong Li, [2004] Issue of August 13;Genome-wide Analysis of ARS (Autonomously Replicating Sequence) Binding Factor 1 (Abf1p)-mediated Transcriptional Regulation in Saccharomyces cerevisiae; *BIOLOGICAL CHEMISTRY* Vol. **279**, No. 33, , pp. 34865–34872,
13. Bonita J. Brewer and Walton L. Fangman, [November 6, 1967]; The Localization of Replication Origins on ARS Plasmids in S. cerevisiae; *Cell*, Vol. **51**, 463-471,
14. http://en.wikipedia.org/wiki/Autonomously_replicating_sequence
15. http://en.wikipedia.org/wiki/Origin_of_replication
16. http://en.wikipedia.org/wiki/Prokaryote
17. http://en.wikipedia.org/wiki/Eukaryote
18. Beginning Perl for Bioinformatics - James Tisdall , Publisher: O'Reilly First Edition October 2001
19. A Practical Guide to Fedora and Red Hat Enterprise Linux, FIFTH EDITION Mark G. Sobell;*PRENTIC HALL*
20. Carol'S. Newlon and James F. Theis,[1993],The structure and function of yeast ARS elements ,*Current Biology* ,**3**:752-758
21. Natalia V. Sernova and Mikhail S. Gelfand ,Identification of replication origins in prokaryotic genomes *Oxford University* ,VOL **9**. NO 5. 376-391

22. Bonita J. Brewer and Walton L. Fangman, November 6, [1967], The Localization of Replication Origins on ARS Plasmids in S.cerevisiae ,*Cell*, Vol**.51**,463-471
23. Marie Touchon, Eduardo P.C. Rocha ,From GC skews to wavelets[2008] : A gentle guide to the analysis of compositional asymmetries in genomicdata *Biochimie* **90** 648e659
24. Vinay Kumar Srivastava and Dharanidhar Dubey,Mapping autonomously replicating sequence elements in a 73-kb region of chromosome II of the fission yeast, Schizosaccharomyces pombe,[August 2007], *Journal of Genetics* Vol. **86**,No.2
25. David M. Gilbert,Evaluating genome-scale approaches to eukaryotic DNA Replication, [OCTOBER 2010] ,*Naturee Reviews GENETICS* VOLUME **11** 673
26. Conrad A. Nieduszynski, Yvonne Knox and Anne D. Donaldson; 2006, Genome-wide identification of replication origins in yeast by comparative genomics; Genes Dev. Cold Spring Harbor Laboratory Press,**20**: 1874-1879
27. Zhijun Wang , Li Jin , Zhenghong Yuan , Grzegorz We˛grzyn , Alicja We˛grzyn; (2009) ,Classification of plasmid vectors using replication origin, selection marker and promoter as criteria; Plasmid *Elsevier* **61** 47–51
28. http://en.wikipedia.org/wiki/Plasmodium_berghei
29. en.wikipedia.org/wiki/Schizosaccharomyces_pombe
30. http://en.wikipedia.org/wiki/Plasmodium_falciparum
31. http://www.perl.com
32. A Beginner's Guide to R,[2009],Alain F. Zuur, Elena N. Ieno, Erik H. W. G. Meesters, *Springer*

Appendix –A
Perl code used for GC skew method

```
#!/usr/bin/perl -w
use strict;
	print "enter file name\n ";
	my $filename=<STDIN>;
	print "\nplease enter the value of window size : ";
	my $wsize=<STDIN>;
	print "\nplease enter the value of displacement : ";
	my $ssize=<STDIN>;
	chomp $filename;
	unless ( open(file1, "$filename") )
	{
		print "Cannot open file \"$filename\"\n\n";
		exit;
	}
	my @file1=<file1>;
	my $file11=join('',@file1);
	if ($file11 !~ /^>/)
		{
			print "not fasta file !!\n";
			exit;
		}
	close file1;
	my $seq='';
	foreach my $line(@file1)
	{
			if($line =~ /^>/)
			{
				next;
			}
		else
			{
			$seq=$seq.$line; # concatenate string
			}
	}
		my @dna=$seq;
		my $dna = join( '', @dna);
		my @dna1 = split( '', $dna );
	 my $l=scalar(@dna1);
	my $countg=0;  my $countc=0;  my $i=0; my $j=0;
	open(fplot1,">$filename.csv");
	while($i<$l-$wsize)
		{
```

```
                $j=$i;
            my $k=$j+$wsize;
            while($j<$k)
                {
                if ($dna1[$j] eq "C")
                    {
                        $countc++;
                    }
                elsif ($dna1[$j] eq "G")
                    {
                        $countg++;
                    }
                $j++;
                }
                my $g=$countg;
                my $c=$countc;
                my $gc=($g-$c)/($g+$c);
                print fplot1 "$i\t$j\t$gc\n";
                $i=$i+$ssize;
                $countg=0;     $countc=0;

        }
close fplot1;

        print"\n";
```

Appendix –B
Perl code used for ARS sequence search

```
#!usr/bin/perl –w
use strict;
print "enter the file name :\n";
my $fname=<STDIN>;
chomp $fname;
print "\nplease enter the pattern to be search : ";
my $read_pat= "WTTTAYRTTTW";
print"\nPlease enter how many mismatch is allowd : ";
my $m =<STDIN>;
chomp $m;
unless(open(fh1, "$fname")){
                print "Cannot open file \"$fname\"\n\n";
                exit;
}
my @fh=<fh1>;
close fh1;
if ($fh[0] !~ /^>/)
                {
                        print "not fasta file\n";
                        exit;
                }
my $seq='';
foreach my $line(@fh)
        {
                        if($line =~ /^>/)
                        {
                        next;
                        }
                        else
                        {
                        $seq=$seq.$line;
                        }
        }
sub trans_pat   #Create string pateern for ARS pattern i.e. subtitute R ,Y ,W
{
  my $pat=shift;
   $pat=~s/R/\[CG\]/g;
```

```
   $pat=~s/W/\[AT\]/g;
   $pat=~s/Y/\[AG\]/g;
  return $pat;
}
open(FH1,">$fname.csv");
sub find_pat
          {
  my ($pat,$seq) = (@_);
   print FH1 "Looking for pattern $pat\n";

          }

find_pat (trans_pat($read_pat),$seq);
my $pat=trans_pat($read_pat);
while ($seq=~m/(?=$pat)/g)
{
print FH1"match at\t$-[0]\t$&\n"     #this will print position  and string of match.
};
foreach my $i (1..(length $read_pat)-($m-1))
{
my $mis_pat = $read_pat;
substr($mis_pat,$i-1,$m)=".{$m}";   #allowing for mismatch of
my $pat1=trans_pat($mis_pat);    #calling of subroutine trans_pat
while ($seq=~m/$pat1/g)

{
print FH1 "match at\t$-[0]\t$&\n";     #Tthis will print position  and string of match.
}
print FH1"$& \n";
}
close FH1;
```

Appendix –C
Other logical code used in project

1. R CODE:- It generate all possible combination of ATGC with repeating their value with desire length.

```
f <-  function(bases, n){apply(expand.grid(rep(list(bases),n)), 1, paste, collapse="")}
#write.table(f(c("A", "T", "C", "G"), n ),file="Myfile.csv",sep=",",row.names=F)
here length is n = 11;  e.g. AAARGCTAGCA
write.table(f(c("A", "T", "C", "G"), 11),file="atgc.csv",sep=",",row.names=F)
```

2. R CODE for ARS pattern graph Plot.

```
# this code plot the whole directory file content into a single graph as bar plot.
library("tcltk2") # include GUI package
library("ggplot2") # include plot package
library("reshape2") # include matrix melting and reshaping package
temp2 <- {}
for (j in 0:3)
{
dir1 <- tk_choose.dir(caption = paste("SELECT DIRECTORY
CONTAINING",j,"MISMATCH FILES")) # GUI directory browser
fns1 <- list.files(path=dir1,pattern="NC*.*csv") # listing all files
fn1 <- substr(basename(fns1), 1, nchar(basename(fns1)) - 4) # finding basename and
deleting .csv extension from each file
mat2 <- data.frame() # empty data frame
for(i in fn1){
file_path <- file.path(dir1,paste(i,".csv",sep="")) # looping through each file and getting
file path
temp1 <- read.csv(file_path, header=FALSE) # reading file to temporary storage
assign(i, temp1) # assigning each file to a different variable
counts1 <- dim(temp1)[1] # taking the row only from dimension
mat1 <- c(assign(paste("counts_",i,sep=""),counts1))  # assigning to no. of rows in each
file
to corresponding named files
mat2 <- append(mat2,mat1)
```

```
  }
  temp2 <- append(temp2,mat2)
 }
CHROMOSOMES <- t(matrix(as.character(fn1))) # reading name of chromosomes to a
matrix
NO_OF_HITS <- matrix(as.numeric(temp2),nrow=4,ncol=length(mat2),byrow=TRUE) #
reading no. of hits to a matrix
df1 <- data.frame(NO_OF_HITS) #making a data frame of above two matrix
colnames(df1)<-fn1
df1$MISMATCH <- factor(c(0:3), levels=c(0:3))
mdf1 <- melt(df1, id.vars="MISMATCH")
colnames(mdf1)<-c("MISMATCH","CHROMOSOMES","NUMBER_OF_HITS")
ylimit1 <- max(mdf1$NUMBER_OF_HITS) + 5000
my_path <- file.path(dirname(dir1),"berg_comparision.png") # writing png file to same
directory
png(paste(my_path), width = 1000, height = 700) # setting png environmennt
ggplot(mdf1, aes(CHROMOSOMES, NUMBER_OF_HITS, fill=MISMATCH)) +
scale_y_continuous(limits = c(0, ylimit1)) + geom_bar(stat="identity",position="dodge",
width=0.5) + geom_text(aes(label=NUMBER_OF_HITS),position =
position_dodge(width=1),angle=90,hjust=-0.1) +
opts(axis.ticks=theme_segment(colour="black",
size=0.5),axis.text.x=theme_text(face="plain",colour="black",size="13",hjust=1,vjust=1,a
ngle=60),axis.text.y=theme_text(face="plain",colour="black",size="13"), title="Different
Mismatch frequency Plots for Plasmodium berghei all chromosome
",plot.title=theme_text(face="bold",size=16))
dev.off()
```