A Self-Adaptive Spectral Rotation Approach to Detection of DNA Sequence Periodicities and Their Relationship with Molecular Mechanisms

by

CHEN Bo

A Thesis Submitted in Partial Fulfilment of the Requirements for the

Degree of Doctor of Philosophy

Department of Industrial and Systems Engineering

The Hong Kong Polytechnic University

June 2011

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

Chen Bo (Name of student)

ABSTRACT

Computational investigations into the relationship and interaction between DNA sequences and cell components help biologists and medical scientists to address many important issues, such as diagnosis of gene-related diseases, medicine development, protein design, and so on. This study initiates a new approach, namely, Self-Adaptive Spectral Rotation (SASR), to investigate the relationship between periodicities in DNA sequences and various molecular mechanisms in cells, including genetic coding and nucleosome formation. This newly developed approach could be very useful in fields of bioinformatics, such as protein-coding region prediction and nucleosome positioning prediction.

Protein-coding region prediction, especially computational methods to find locations of protein-coding regions in uncharacterized DNA sequences, is a meaningful issue in computational molecular biology. In this study, the SASR approach is first developed to visualize a coding related feature, i.e., the Triplet Periodicity (TP) or 3bp (base pairs) periodicity, in DNA sequences. Applications on real genomic datasets show that, in SASR's output, the graphic patterns for coding and non-coding regions differ so significantly that the former can be visually distinguished from the latter. Such visualization by the SASR approach requires no training process, and takes the advantage of "auto-scale analysis ability" from human vision. However, as a visualization method, the SASR approach does not provide exact numerical predictions. Therefore, a T-Z-T approach is developed to extract

numerical information from the SASR's graphic result. The combination of the SASR and the T-Z-T provides computational predictions of coding regions without any training process. Moreover, the predictions from this SASR based approach are more robust than those from commonly used methods based on Hidden Markov Model (HMM), since this new approach is not sensitive to input errors contained in DNA sequences.

Experimental studies on nucleosome positioning have revealed the preference of nucleosome binding for certain regions of a DNA sequence. However, it is still not clear whether or not such a binding preference is sequence-specific. Therefore, the study on the relationship between sequence features and nucleosome formation is of great significance. A major concern in this issue is the ~10bp periodicity property in DNA sequences, which appears to be associated with the structure of DNA helixes and the formation of nucleosomes. In this study, the original SASR approach is extended to investigate the relationship between nucleosome formation and the ~10bp periodicity of dinucleotides in DNA sequences. A Genetic Algorithm (GA) based method is developed to identify which dinucleotide combination mostly connects its ~10bp periodicity with nucleosome formation. The results from the GA support the "sequence-specific" argument of nucleosome formation. Meanwhile, they also suggest that some dinucleotides connect their ~10bp periodicity with nucleosome formation only in some local regions. Moreover, the ~10bp periodicity of dinucleotides is associated with not only the occurrence of nucleosome formation, but also the binding preference for the phase in the ~10bp period.

Besides the TP and the ~10bp periodicity, some other unknown periodicity

properties may also be contained in DNA sequences, and may have some connections with some important molecular mechanisms. Investigations of new periodicity properties might help with the computational studies of sequence-specific molecular mechanisms in organisms. In this study, another extension of the SASR approach, i.e., the mature SASR, shows its ability to detect a hypothetical anti-TP property in DNA sequences. Some real DNA fragments are found with such an anti-TP property by using the mature SASR. However, the universality of this property in genomes and its biological interpretation need further investigations.

ACKNOWLEDGMENTS

This research would not have been a success without the help of many people who give me the greatest support.

I would like to acknowledge my supervisor, Dr. P. Ji, in the Department of Industrial and Systems Engineering of The Hong Kong Polytechnic University, for his enlightening and patient guidance throughout my research project. His consistent encouragement, constructive suggestions, and countless discussions have helped me to expedite the development of my research project. His deep understanding and profound knowledge have broadened my view in this field of study.

Besides, my sincere gratitude goes to Professor Xu in Fuzhou University for his careful proofreading which enables me to present such a linguistically polished dissertation. I also appreciate my friends and colleagues for their tremendous support and encouragement. Meanwhile, I am definitely indebted to The Hong Kong Polytechnic University whose financial support renders this research project possible.

Appreciation should also be given to my beloved parents. As my closest relatives and friends, they have always been concerned about me and look forward to any achievement made in my academic pursuit. It is hard to imagine what progress I could have made in my life and career without this selfless and unconditional parental love and care. Words fail to express my deepest gratitude for them. In particular, this dissertation is in everlasting memory of my dearest mother who passed away when I was composing the manuscript. In the public eyes, she was an outstanding engineer, a filial daughter, a virtuous wife, and a loving mother. Her sudden departure left us in great sorrow and eternal regret. May this thesis console her soul in heaven.

Finally, I would like to acknowledge all parties involved from the bottom of my heart, including those mentioned above together with those who helped me indeed but I missed to thank previously.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

A	Adenine
Ac	Accuracy
ACF	Autocorrelation Function
BLAST	Basic Local Alignment Search Tool
bp	Base Pairs
C	Cytosine
CDF	Cumulative Distribution Function
C. elegans	*Caenorhabditis elegans*
CGR	Chaos Game Representation
ChIP	Chromatin Immunoprecipitation
C strand	Crick strand
DNA	Deoxyribonucleic Acid
DP	Dynamic Programming
EST	Expressed Sequence Tag
FCGR	Chaos Game Representation with Frequencies
FSND	Fixed Scale Numerical Differentiation
G	Guanine
GA	Genetic Algorithm

GHMM	Generalized Hidden Markov Model
HMM	Hidden Markov Model
H set	Head set
LCS	Left Cyclic Shift
LRC	Long-Range Correlation
LRD	Long-Range Dependence
MNase	Micrococcal Nuclease
mRNA	Message Ribonucleic Acid
M set	Middle set
MSP	Maximal Segment Pair
mtDNA	Mitochondrial Deoxyribonucleic Acid
NCBI	National Center for Biotechnology Information
NFR	Nucleosome-Free Region
ORF	Opening Reading Frame
OSCM	Optimized Spectral Content Measure
PDF	Probability Density Function
PP	Purine-Pyrimidine
Pr	Precision
PSD	Power Spectral Density
RCS	Right Cyclic Shift
RNA	Ribonucleic Acid

RR	Rightward Rate
rRNA	Ribosomes Ribonucleic Acid
SASR	Self-Adaptive Spectral Rotation
S. cerevisiae	*Saccharomyces cerevisiae*
SCM	Spectral Content Measure
Sn	Sensitivity
Sp	Specificity
S. pombe	*Schizosaccharomyces pombe*
SRM	Spectral Rotation Measure
SRR	Symmetrical Rightward Rate
STFT	Short Time Fourier Transform
T	Thymine
TP	Triplet Periodicity
TPM	Triplet Periodicity Matrix
tRNA	Transfer Ribonucleic Acid
U	Uracil
UCSC	University of California Santa Cruz
WS	Weak-Strong
W strand	Watson strand
τ-PM	τ-Periodicity Matrix

CHAPTER 1

INTRODUCTION

Nowadays, owing to the great development of DNA sequencing techniques and bioinformatics, one can easily and conveniently get DNA sequences as well as feature annotations of various organisms by accessing public databases such as GenBank. It provides opportunities to computationally investigate the relationship and interaction between DNA sequences and components of cells. And such computational investigations can help biologists and medical scientists in addressing some significant issues, including diagnosis of gene-related diseases, medicine development, protein design, and so on. This study focuses on the relationship between periodicities in DNA sequences and various molecular mechanisms in cells, including genetic coding and nucleosome formation, and help with some significant issues, such as protein-coding region prediction and nucleosome positioning prediction.

1.1 Protein-coding region prediction and Triplet Periodicity (TP)

A significant function of DNA is to instruct the synthesis of proteins, which are basic organic compounds made of amino acids arranged in a linear chain. However, a DNA sequence is roughly divided into coding and non-coding regions, and only the

coding regions serve as instructors of protein synthesis. Therefore, discriminating coding and non-coding regions and computationally locating coding regions in uncharacterized DNA sequences becomes a significant issue in DNA research (Fickett, 1996), which is called coding region prediction.

This study attempts to develop a computational approach for coding region prediction based on a Triplet Periodicity (TP) property, which was first considered as a coding related feature by Fickett (1982). In Fickett's work (Fickett, 1982), for each possible separation k in DNA fragments, he counted the number of times two Thymine (T) appearing with k nucleotides between them, and compared this with the count expected in a model where bases were chosen independently. He discovered that in the coding fragments, the difference between the practical count and the expected count is big when k = 2, 5, 8, …, (2+3n), but in the non-coding fragments, there is no such regularity (Figure 1.1). Accordingly, he claimed that it is a simple and universal difference between coding and non-coding regions.

After Fickett's first description, this property is then commonly described involving the term of Fourier analysis. It is said that when DNA fragments are represented in some digital ways, the Fourier spectrums of the coding fragments have a peak at the frequency $k = N/3$ (N stands for the length of the region), while there is no such peak for the non-coding fragments. This phenomenon, which is called Triplet Periodicity (TP), comes from the fact that coding regions consist of triplets (codons),

and the usage of codons for coding amino acids are highly nonrandom. Analysis of the TP is said to be quite distinctive for identifying coding regions in DNA sequences.

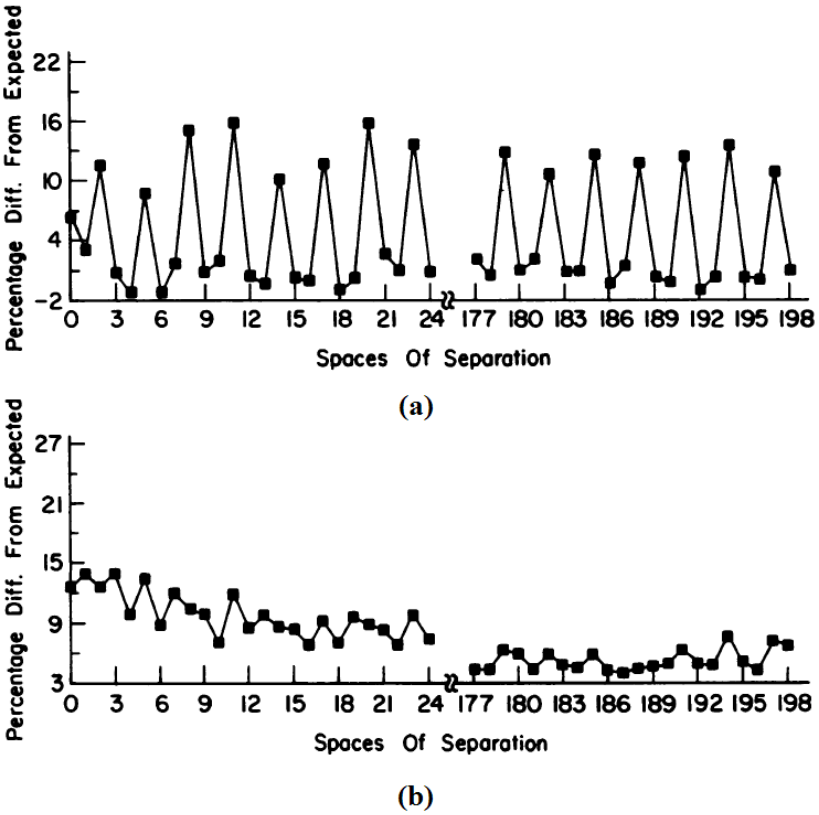


Figure 1.1 Fickett's autocorrelation graphs (Fickett, 1982) (a) coding fragments (b) non-coding fragments

1.2 Nucleosome formation and ~10bp periodicity

Nucleosomes are basic structural repeating units of chromatin, each consisting of a histone core around which the DNA double helix is wrapped (Luger et al., 1997; Segal et al., 2006; Nair, 2009; Jiang and Pugh, 2009). Nucleosomes are ubiquitous in the chromosome, but the principle of nucleosome formation is not clear nowadays. For long, nucleosomes had been thought to be formed randomly along the DNA

sequence, but recent studies have thrown light into the non-random nucleosome positioning on the chromosome and its function in regulating genomic functions (Kornberg and Lorch, 1999; Wyrick et al., 1999; Cooper, 2000; Segal et al., 2006).

It has been debated whether nucleosomes are formed under regulation of the cell or following instructions of the DNA sequence (Nair, 2009). A ~10bp (base pairs) periodicity of dinucleotides discovered in high nucleosome formatting regions of DNA sequences seems to support the later argument. The ~10bp periodicity of dinucleotides was first noted by Trifonov (Trifonov and Sussman, 1980; Trifonov, 1980). Trifonov calculated the summation of the autocorrelation functions for all 16 dinucleotides of the SV40 DNA sequence. Besides the obvious TP property, he discovered another weak repeating with a ~10bp period (Figure 1.2a). He further suggested that nucleosomes are formed by sequence-dependent bending of the DNA molecule (Trifonov and Sussman, 1980) and the unidirectional bending of a DNA molecule is ascribed to regular insertions of some “wedges”, consisting of two nonparallel adjacent base pairs (Figure 1.2b).

After that, many other researches have confirmed this periodicity property associated with nucleosome formation along the DNA sequence (Davey et al., 1995; Hosid et al., 2004; Albert et al., 2007; Mavrich et al., 2008; Chen et al., 2008). Meanwhile, various explanations are raised for the relationship between the ~10bp periodicity and nucleosome formation. However, there are still insufficient solid

evidences to prove that the DNA sequence “encodes” nucleosome formation.

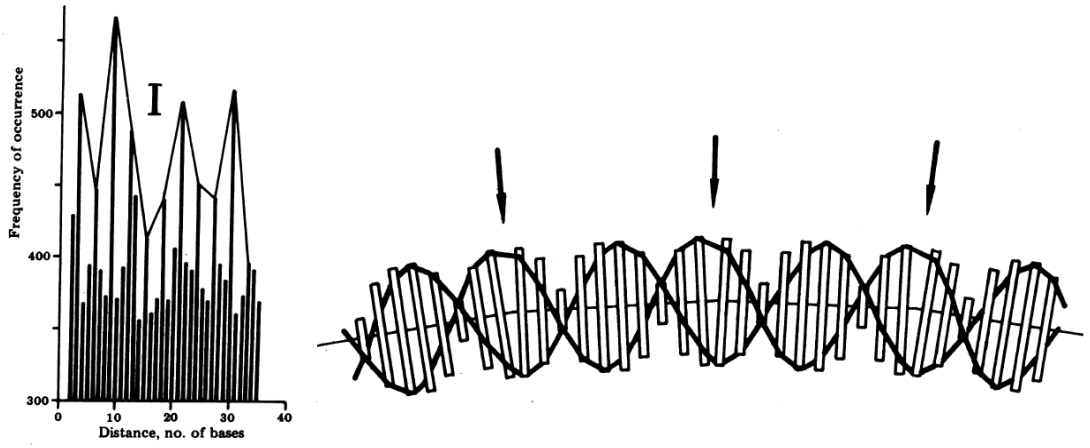


Figure 1.2 Trifonov’s finding of the ~10bp periodicity and his explanation for the bending of a DNA molecule (Trifonov and Sussman, 1980) (a) The summation of the autocorrelation functions for all 16 dinucleotides of the SV40 DNA sequence (b) The unidirectional bending of a DNA molecule by regular insertions of nonparallel adjacent base pairs (arrows)

1.3 Research objectives

The aim of this research is to study on molecular mechanisms related to sequence periodicities. Considering the two well-known periodicity-related molecular mechanisms, i.e., genetic coding and nucleosome formation, this aim is further presented in the following detailed objectives:

(1) To find a suitable approach to extract and represent periodicity properties hidden in DNA sequences, without any training or previous knowledge. By using the proposed approach, one can easily access periodicity properties in

any region of the sequence and with any analysis scale;

(2) By using the proposed approach in (1) to extract the TP property in DNA sequences, and develop a TP-based computational method for protein-coding region prediction, without any training process;

(3) By using the proposed approach in (1) to extract the ~10bp periodicity property in DNA sequences, and study the relationship between the ~10bp periodicity in DNA sequences and the nucleosome binding preference. Try to find out evidences to support the argument that the binding preference is (or is not) sequence-specific;

1.4 Research outline

Bioinformatics is the application of computer science and information technology to the field of biology and medicine. The outline of this research work is presented here. First, to resolve the current problem in the field of bioinformatics, a new approach is proposed from the theoretical basis after the literature review. Operable algorithms are developed so that the new approach could be implemented in practice and the computational results obtained from the approach are compatible with those from other contemporary methods. Then theoretical proofs are given to demonstrate the efficiency of the approach in dealing with the subject. Besides, the efficiency is verified by applications to some simulated datasets and real DNA

datasets collected from public databases. Some other popular methods for the same problem are also applied to the same datasets. From the comparisons with such popular methods, the advantages and the limitations of the new approach are revealed. After that, the approach is further extended to deal with other related problems.

The source data of the DNA sequences are collected from the NCBI's Entrez Nucleotide database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide). The GenBank reference number of each DNA sequence used in this work has been given when it first presents in this thesis. Record files used in this work for the nucleosome positioning problem are downloaded from the UCSC Genome Bioinformatics Site (http://moma.ki.au.dk/genome-mirror/).

1.5 Thesis scope

The remaining chapters of this thesis are organized as follows.

Chapter 2 gives a comprehensive literature review on protein-coding region prediction and nucleosome formation. It includes a brief introduction to DNA and DNA sequence. Some necessary knowledge about gene and genetic coding is presented, as well as current methods for computational coding region prediction. Besides, a brief introduction to nucleosome is given, including nucleosome's structure and its role in cells. Main findings and arguments about nucleosome

formation are also reviewed in this chapter.

Chapter 3 presents a new approach to visualize the TP property in DNA sequences. The principle and the algorithm of this approach are demonstrated in detail. After that, some computational analyses on its graphic output are proposed, so that numerical results can be extracted and computational predictions of coding regions can be provided.

As verifications of the new approach proposed in Chapter 3, results of various applications are given in Chapter 4. This chapter tests the computational complexity of the new approach, shows its practical behaviors when it is applied to real DNA sequences, and evaluates the performance of the computational coding region prediction by using this approach. The results are compared with those obtained by other methods, so that the advantages of the new approach are revealed.

In Chapter 5, the new approach is further developed for applications in some other fields. It is first extended for the investigation of the periodicity property with any rational period, so that it can help with the studies of the ~10bp periodicity and nucleosome formation. Besides, another extension of the new approach is developed to detect a hypothetical anti-TP property, which probably exists in real DNA sequences.

Finally, conclusions of this study and prospects for future work are given in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

In this chapter, some previous researches will be reviewed, including the background and knowledge of relevant fields in molecular biology, previous methods for computational coding region prediction, as well as debated issues in nucleosome positioning. Such a literature review is presented for a better understanding of the rest parts of this thesis.

2.1 DNA and DNA sequence

DNA is known as the carrier of genetic information in nearly all living beings. Studies on DNA structure and DNA behaviors are crucial in genetics and molecular biology.

2.1.1 A brief introduction to DNA

DNA, a short term for Deoxyribonucleic Acid, is a nucleic acid that contains genetic information in all known living organisms and some viruses. It restores genetic information for a long term and instructs the construction of other components of cells, such as proteins and RNA molecules.

In 1869, DNA was firstly isolated by a Swiss physician, Friedrich Miescher

(Dahm, 2008). And then researchers elaborated on finding a correct structure model of DNA. Levene (1919) suggested that DNA is a string of nucleotide units linked together through phosphate groups and indicated that a nucleotide unit consists of base (Adenine, Thymine, Guanine, or Cytosine), sugar, and phosphate (Figure 2.1). This structure is similar to what is accepted today as a single strand of DNA. Nowadays, the structures of bases, sugar, phosphate ion, and nucleotide unit in DNA have been approved as in Figure 2.2. Nucleotide units, which consist of base, sugar, and phosphate, are joined together in a linear manner through phosphate groups attached to the 3' and 5' positions of the neighboring sugars as shown in Figure 2.3 (Neidle, 2008). Hence, a full strand of DNA has two ends with sugars, whose 3' or 5' position is not attached to any other nucleotide, and they are commonly called the 3' end and the 5' end.

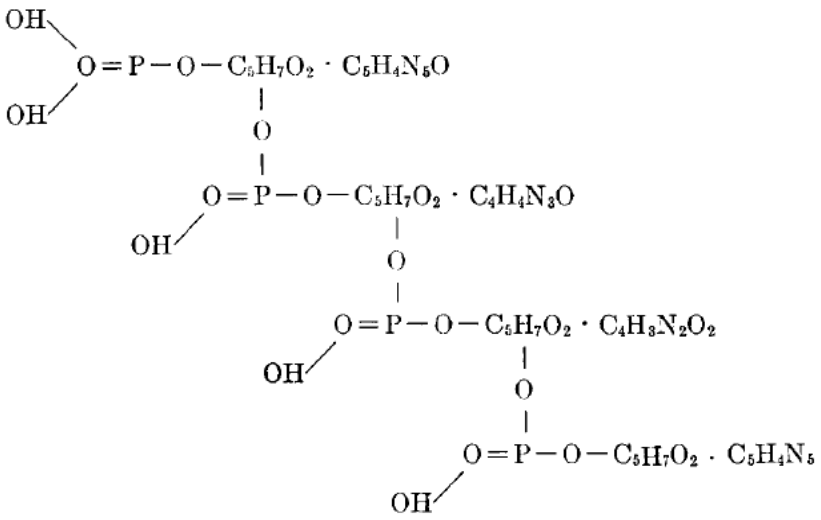


Figure 2.1 DNA structure in Levene's work (1919)

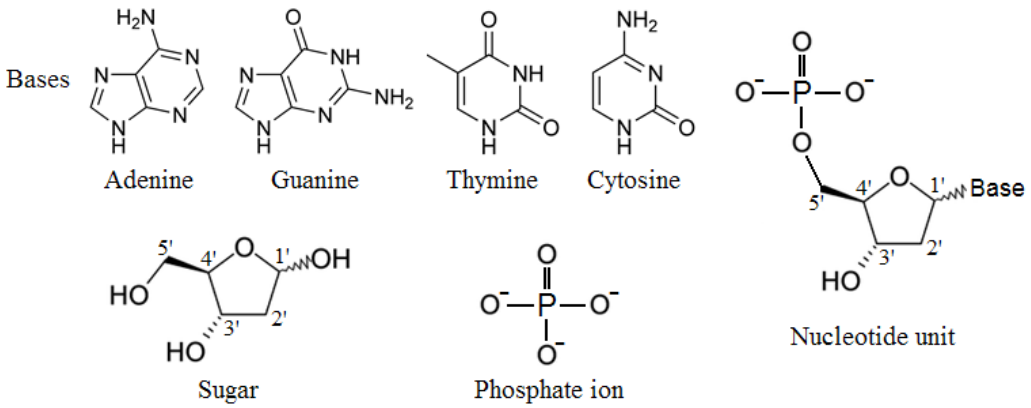


Figure 2.2 Structures of bases, sugar, phosphate ion, and nucleotide unit

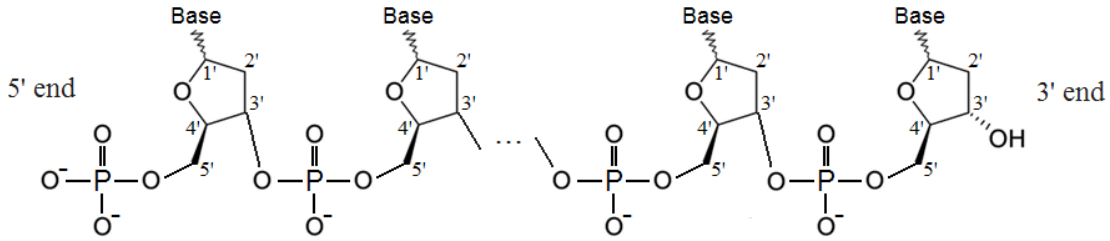


Figure 2.3 A full strand of DNA with two ends

However, the complete structure of DNA is far more complicated than a single strand. In the early 1950s, Erwin Chargaff et al. (Chargaff, 1950; Chargaff et al., 1951 and 1952) found that in any DNA molecule, the amount of Adenine (A) equals to that of Thymine (T) and the amount of Guanine (G) equals to that of Cytosine (C). This finding was then developed and known as the "Chargaff Rules". In 1952, Franklin and Gosling got an X-ray diffraction image of DNA (Franklin and Gosling, 1953). Based on Chargaff's finding and Franklin and Gosling's image (Figure 2.4a), Watson and Crick (1953) suggested a double-helix model which is now accepted as

the first correct DNA structure model. In Watson and Crick's model, a DNA molecule has two helical strands, each coiled around the same axis (Figure 2.4b). Both strands follow right-handed helices, but the sequences of the atoms in the two strands run in opposite directions (sequences are commonly read from the 5' end to the 3' end). The bases are on the inside of the helix and the phosphates on the outside. The bases from the two strands are connected via hydrogen bonds and form base pairs to hold the DNA double helix together.

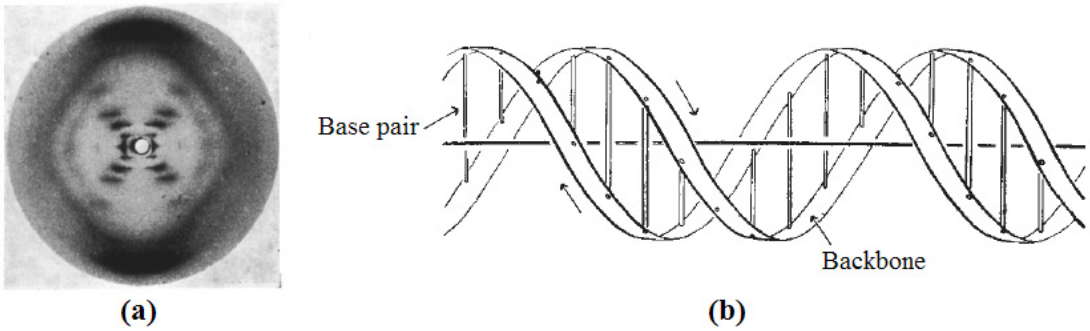


Figure 2.4 Discovery of the DNA structure model (a) Franklin and Gosling's X-ray diffraction image (1953) (b) Watson and Crick's DNA model (1953)

In Watson and Crick's base pairing principle (Watson and Crick, 1953), Adenine (A) forms a base pair with Thymine (T), so does Guanine (G) with Cytosine (C). This principle explains why the amount of Adenine equals to that of Thymine and the amount of Guanine equals to that of Cytosine in any DNA molecule, and it also implies that the base sequences of the two strands determine each other. The chemical structures of both the G-C pair and the A-T pair are shown in Figure 2.5. The initial pairing model assumed just two hydrogen bonds for the G-C pair and the

third hydrogen bond was found three years later (Wain-Hobson, 2006; Neidle, 2008).

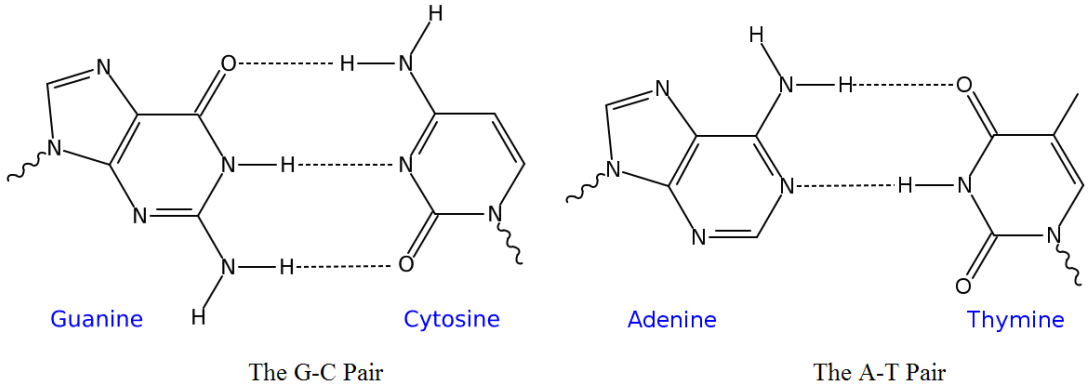


Figure 2.5 Base pairing. The dot lines stand for hydrogen bonds.

2.1.2 DNA sequence

DNA sequence is an abstract model, by which we focus on significant factors and ignore irrelevant factors in the research. Since the structure of the DNA backbone (sugar and phosphate) is fixed, it is believed that DNA carries the genetic information by its unique sequence of base pairs (Alberts et al., 1994). Therefore, in most cases, only the sequence of base pairs is concerned instead of its complicated double-helix structure. Moreover, as mentioned previously, the base sequences of the two strands determine each other according to Watson and Crick's base pairing principle, so it is only required to focus on one of the two base sequences. Therefore, in most cases, a meaningful representation of a DNA molecule is to represent the base sequence of its primary strand, namely, the W strand (the other one is called the C strand). In common contexts, the base sequence of the W strand is called a "DNA

sequence", or "sequence" for short.

String representations

A straightforward and most commonly used representation of a DNA sequence is a character string, which directly points out the base sequence of the W strand from the 5' end to the 3' end (it is the default direction in the following sections). In this string representation, the bases Adenine, Thymine, Guanine, and Cytosine are denoted as 4 characters, i.e., A, T, G, and C, respectively (Alberts et al., 1994). It has been logically accepted and used as the default representation for a long time since the research focus is on the base sequence. However, there are still many other representations, which provide different views for DNA sequences.

One significant representation is a binary string representation first advanced by Voss (1992). In this representation, a DNA sequence is represented as four binary strings, i.e., $u_A(t)$, $u_T(t)$, $u_G(t)$, and $u_C(t)$. $u_\Lambda(t) = 1$ if and only if base Λ (A, T, G, or C) appears at position t in the DNA sequence S. An example is shown below:

t: 123456789
S: *ATGATGACG*
$u_A(t)$: 100100100
$u_T(t)$: 010010000
$u_G(t)$: 001001001
$u_C(t)$: 000000010

This binary string representation is then comprehensively used in some computational analyses of DNA sequences (Tiwari et al., 1997; Anastassiou, 2000;

Kotlar and Lavner, 2003; Datta and Asif, 2004; Eftestol et al., 2006).

Epps et al. (2008) presented a general representation model for sequences with an alphabet size of M. In this model, a sequence $S = \{s[1], s[2], \ldots, s[N]\}$ (N is the length of S) is represented as:

$$s[t] = \sum_{m=1}^{M} a_m b_m[t]$$

Here each a_m is an integer assigned to the m^{th} character in the alphabet and each $b_m[t]$ is a binary sequence indicating whether the m^{th} character appears at position t in the sequence. For a DNA sequence, $M = 4$ and $b_m[t]$ here obviously equals to $u_\Lambda(t)$ described in Voss's representation. It shows that the binary strings, i.e., $u_\Lambda(t)$ or $b_m[t]$, are the essential data for a certain sequence in Epps's model and the coefficient a_m maps a base (A, T, C, or G) into an integer for computational analyses.

Graphic representations

Compared with the string representations mentioned above, graphic representations provide visual patterns for DNA analyses. Most of such graphic representations can be classified as random walks. The concept of random walk was first introduced by Pearson (1905). It is central to probability theory and still occupies the mathematical mind today (Stewart, 2001). A random walk can be considered as a mathematical formalization of a trajectory that consists of taking successive steps. The walk trace can be in a certain space such as the one-

dimensional space, the two-dimensional space, and so on. And the steps the walk takes can be irregularly random or follow instructions from a certain sequence or a generating algorithm.

A DNA walk is a random walk following instructions from a DNA sequence. It is said that trends or correlations in nucleotide composition of a DNA sequence can be shown in terms of different classes of DNA walks (Cebrat and Dudek, 1998). For example, the Purine-Pyrimidine (PP) DNA walk is a random walk in the one-dimensional space and it shows the difference in purine versus pyrimidine composition. In the PP walk, the DNA sequence is read base by base. The walker goes "up" if it finds a purine (G or A) and goes "down" if it finds a pyrimidine (C or T). Figure 2.6a is an example of the PP walk described in Cebrat and Dudek's work (Cebrat and Dudek, 1998). Different from the PP walk, Berthelsen's GC-AT DNA walk (Berthelsen et al., 1992) in the two-dimensional space shows the bias of the bases in a DNA sequence. In such a walk, the walker goes "up", "down", "left", or "right" when it meets "G", "C", "A", or "T", respectively. Figure 2.6b shows an example of the GC-AT walk from Berthelsen's work (Berthelsen et al., 1992).

Another significant DNA walk is the Z curve proposed by Zhang and Zhang (1994). The Z curve is said to be a curve in 3-D constituting a unique representation of a given DNA sequence and the DNA sequence can be reconstructed from the curve. In this presentation, the curve is also generated by a series of steps. For each

step t, count the cumulative numbers of the four bases appearing in the subsequence from the first to the t^{th} base and denote the cumulative appearing numbers of A, T, G, and C as A_t, T_t, G_t, and C_t, respectively. Then the Z curve is defined as a point set $\{(x_t, y_t, z_t) \mid t = 1, 2, \ldots, N\}$, where N is the length of the sequence, and for each step t, the point (x_t, y_t, z_t) is calculated by:

$$\begin{cases} x_t = 2(A_t + G_t) - t \\ y_t = 2(A_t + C_t) - t \\ z_t = 2(A_t + T_t) - t \end{cases}$$

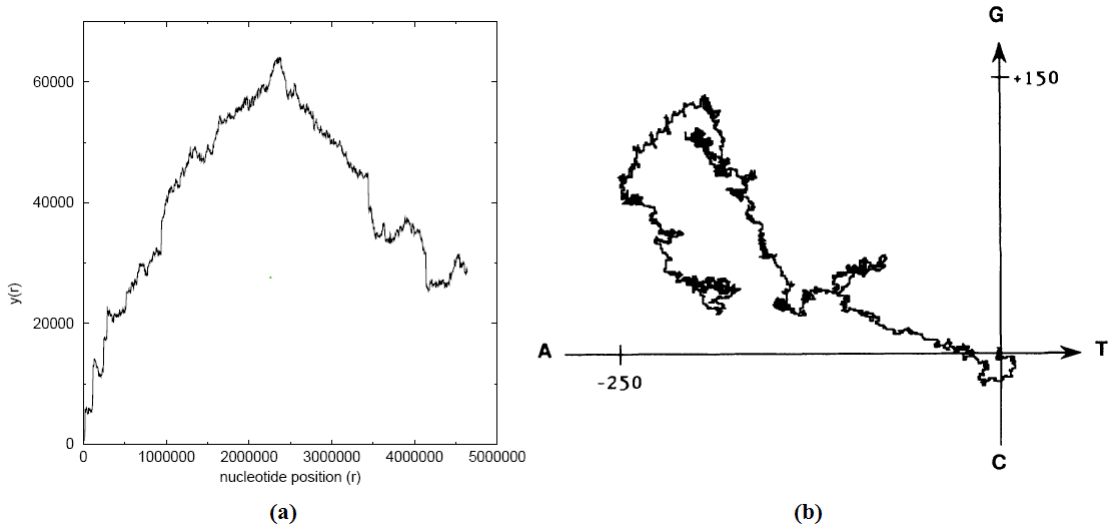


Figure 2.6 Examples of the DNA walks (a) The PP DNA walk in Cebrat and Dudek's work (1998) (b) The GC-AT DNA walk in Berthelsen's work (1992)

Figure 2.7 shows some Z curves collected from Zhang et al.'s work (Zhang et al., 2003). Currently, a database containing Z curves for various organisms has been developed (Zhang et al., 2003), and meanwhile, such a representation has been used in researches on many topics, including replication origin identification (Zhang and

Zhang, 2002; Zhang and Zhang, 2005), gene prediction (Yan et al., 1998; Zhang and Wang, 2000; Guo et al., 2003), comparative genomics (Zhang and Zhang, 2003), and so on.

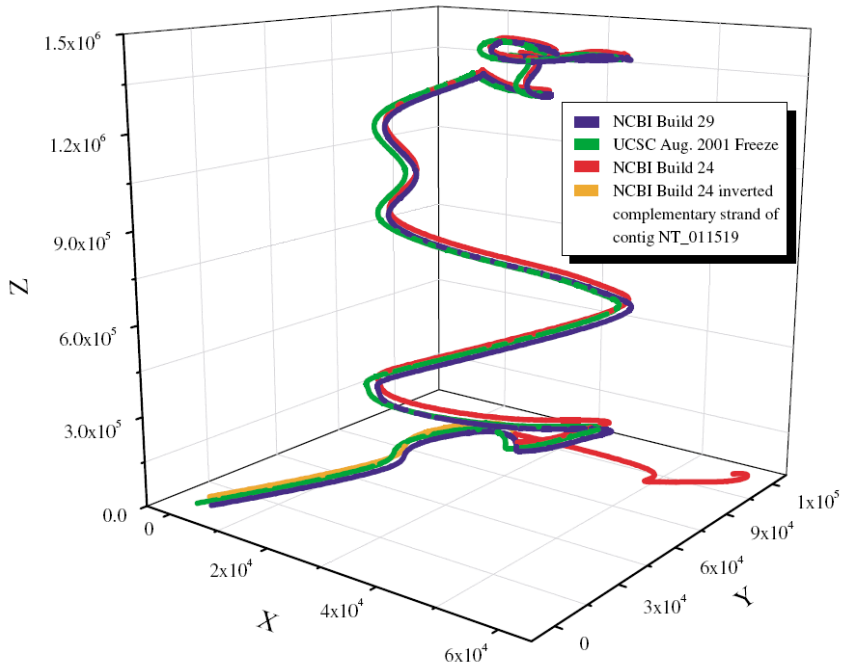


Figure 2.7 The Z curves for the human chromosome 22 based on different assemblies (Zhang et al., 2003)

Besides the DNA walks, there are some other graphic representations, including Chaos Game Representation (CGR), which was originally proposed by Jeffrey (1990). In this representation, an image with special fractal patterns is created by an iterative process to represent a DNA sequence. The iterative process can be briefly introduced as follows: Initially, we have a blank square with four vertices, each of which denotes a base, i.e., “A”, “T”, “G”, or “C”, and a point is put in the centre of

the square. The DNA sequence is read base after base and the point moves step by step correspondingly until the end of the sequence. For each step, the point moves forward to the vertex corresponding to the base read from the sequence and leaves a dot on the image. The step length is half the distance between the last position and the vertex. After this iterative process, we can get a CGR image which is a set of dots left by the movements of the point (Figure 2.8).

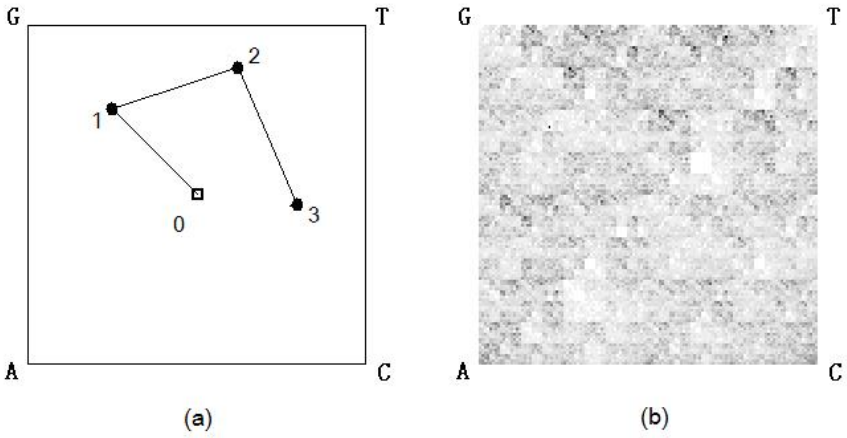


Figure 2.8 CGR images (a) The first 3 steps in generating the CGR of the sequence "GTC…" (b) A complete CGR image

Jeffrey (1990) pointed out that the DNA sequence of certain species has its unique fractal pattern in its CGR image. Since it was originally put forward for visualized comparison, he did not give further mathematical descriptions. Researchers tried to describe CGR by using different theories. Oliver and Bernaola (1993) analyzed CGR images by using informatics theory and mainly discussed the entropies of sub-sequences and the non-random features of sequences. Then CGR

was considered to be a representation of status. In 1998, Baldi and Brunak (1998) pointed out that CGR is essentially a Markov probability table and the CGR space is a reference system that every probable sequence has a position in it. Accordingly, all probable sequence patterns can be represented in this space.

There are a number of modifications of CGR, which provide some significant properties and extend the applications. Randić et al. developed, from CGR, a unique graphical representation of protein sequences (Figure 2.9a) by assigning the four bases into symmetrical positions in 3-D (Randić et al., 2004). In 2008, Randić further refined CGR by mapping the 2-D CGR to a spectrum and conducted some discussions (Randić, 2008).

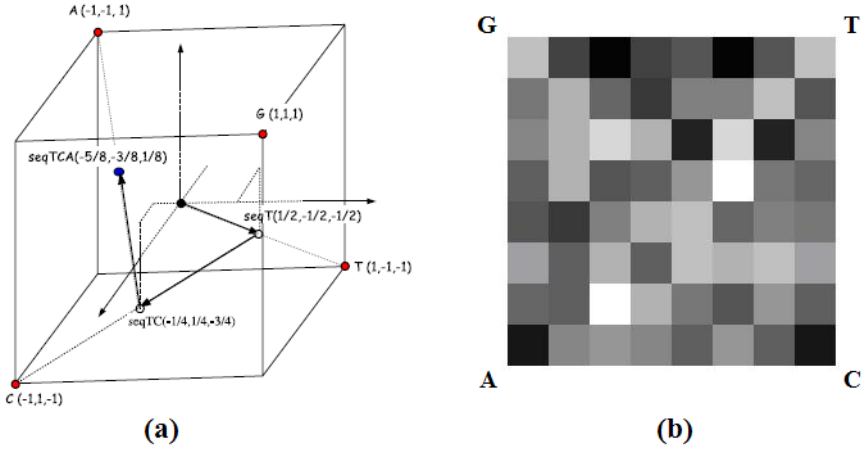


Figure 2.9 CGR modifications (a) The unique graphical representation of sequence "TCA…" in Randić's work (2004) (b) The FCGR image

A most significant modification of CGR is FCGR (Chaos Game Representation with Frequencies). Deschavanne (1999) modified CGR to represent frequencies of

sub-sequence patterns and Almeida et al. (2001) proposed the name “FCGR” for this modification. In FCGR, the original CGR image is divided into 4^L small squares and each square is filled with a color related to the number of dots in it (Figure 2.9b). It is pointed out that each square represents a pattern with a length of L and the color of this square represents the frequency (times) of the pattern in the complete sequence. Such an FCGR has mainly been used to analyze the distance and correlation between DNA sequences in some researches (Yu et al., 2004; Wang et al., 2005).

2.2 Genes and protein synthesis

As an objective of this study, an attempt is made to develop a computational method for coding region prediction. Therefore, it is necessary to review the background of genes and protein synthesis.

2.2.1 Brief background of genetics

The science of genetics dates back to 1900, since three botanists rediscovered Gregor Mendel’s work (Alain and Floyd, 1993) which first suggested the existence of genes. Mendel started his pea hybridization experiments in 1856 and for 8 consecutive years he grew thousand of peas for the research. He finished his experiments in 1863 and in 1866 presented his report paper, which is now considered as the founding document of science of genetics, since it provided the first theory of

heredity based on real experimental evidence and contained the first two fundamental laws of heredity: the law of segregation of genes and the law of independent assortment of genes (Alain and Floyd, 1993).

After the rediscovery of Mendel's work, the science of genetics was developed very rapidly. In 1928, Griffith reported his experiment, nowadays known as the Griffith's experiment (Griffith, 1928). He tested the behaviors of two strains of *Pneumococcus* (which infects mice), i.e., type III-S (smooth) and type II-R (rough) strain. The smooth strain covers itself with a polysaccharide capsule that protects it from the host's immune system, while the rough strain doesn't have that protective capsule and is defeated by the host's immune system. It means that the smooth strain can kill the host (mouse), but the rough strain is harmless. In his experiment, bacteria from the smooth strain were killed by heat (turns to harmless) and mixed into the rough strain bacteria. While neither alone harmed the mouse, the combination killed the mouse (Figure 2.10). Griffith explained that it is because of a "transforming principle", by which the rough strain bacteria is "transformed" into smooth strain bacteria following some instructions from part of the dead smooth strain bacteria. And then the Avery-MacLeod-McCarty experiment, extended from the Griffith's experiment, indicated that DNA is the substance that causes bacterial transformation (Avery et al., 1944). DNA is then deemed to contain genetic information.

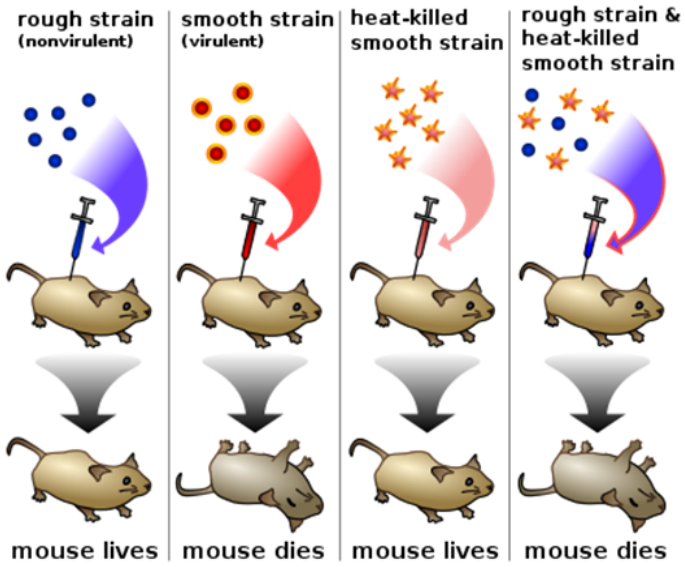


Figure 2.10 Sketch of the Griffith's experiment

Beadle and Tatum (1941) discovered that mutations in some fragments in the DNA sequence (genes) could cause errors in metabolic pathways and they stated a "one gene, one enzyme" view, which later became "one gene, one polypeptide". Gene was then considered as a blueprint of a protein (Gerstein et al., 2007). This view, together with Avery et al.'s finding (Avery et al., 1944) and the Watson and Crick's DNA model (Watson and Crick, 1953), established the "central dogma", which is a basic framework for the transfers among DNA, RNA (see Section 2.2.2), and proteins in living organisms. However, there is an increasing realization that genes are not only the instructors of protein synthesis and some RNA components, but also have various regulating functions in cells (Mattick, 1994; Frank and Pace, 1998; Doherty and Doudna, 2001; Doudna and Cech, 2002; Barciszewski and Erdmann, 2003). Although the concept of gene has been developed and extended

from the “one gene, one polypeptide”, the “central dogma” can be still considered as a correct principle by which a DNA molecule instructs the synthesis of proteins in the cell.

2.2.2 Central dogma and genetic coding

The central dogma was first presented by Crick (1958). The general transfer processes mentioned in this dogma include transcription, translation, and DNA replication (Figure 2.11).

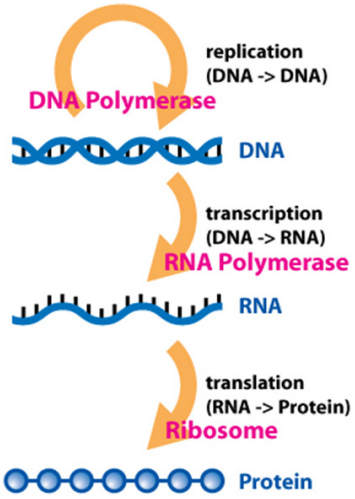


Figure 2.11 The general transfer processes mentioned in central dogma

Transcription is the synthesis of RNA molecule under instructions of certain DNA fragments, i.e., the genes (Alberts et al., 1994). Ribonucleic Acid (RNA) has a single strand structure, which is similar to the single strand in DNA. But RNA contains ribose sugar instead of deoxyribose sugar in DNA and RNA has base Uracil (U) rather than Thymine (T) in DNA. In transcription, the two strands of a DNA

fragment separate progressively, and meanwhile, one strand (W or C) of the fragment serves as a template, so that a complementary strand of RNA molecule can be synthesized by the base pairing principle (A-U and G-C). After that, the double helix re-forms and the RNA molecule is released.

RNA transcripts which direct the synthesis of proteins are called message RNA (mRNA), while other RNA transcripts serve as some other functional roles, such as transfer RNA (tRNA), ribosomes RNA (rRNA) (Alberts et al., 1994), or some regulation units in the cell (Wang et al., 1996; Liu et al., 1997; Bussemakers et al., 1999; Doherty and Doudna 2001; Barciszewski and Erdmann, 2003). Moreover, it is found that there is another process named "RNA splicing", after transcription. In RNA splicing, some parts in an mRNA molecule, namely, introns, are removed, and the rest parts, namely, exons, join together to form a spliced mRNA molecule, which finally direct the synthesis of proteins (Berget et al., 1977; Alberts et al., 1994). Such an RNA splicing process is optional and it mostly happens to eukaryotic RNA molecules.

Translation only happens to mRNA. In this process, the bases in an mRNA molecule are read in a serial order in sets of three (triplets). Each triplet of bases, namely, codon, specifies an amino acid and the mapping from codons to amino acids, namely, the genetic code, is shown in Table 2.1. The amino acids are joined linearly in the same order, in which their corresponding codons appear in the mRNA

molecule, and a protein (a polypeptide) is then synthesized (Alberts et al., 1994). Generally, as shown in Figure 2.12, by transcription, splicing (if any), and translation, proteins are synthesized under instructions of a DNA blueprint.

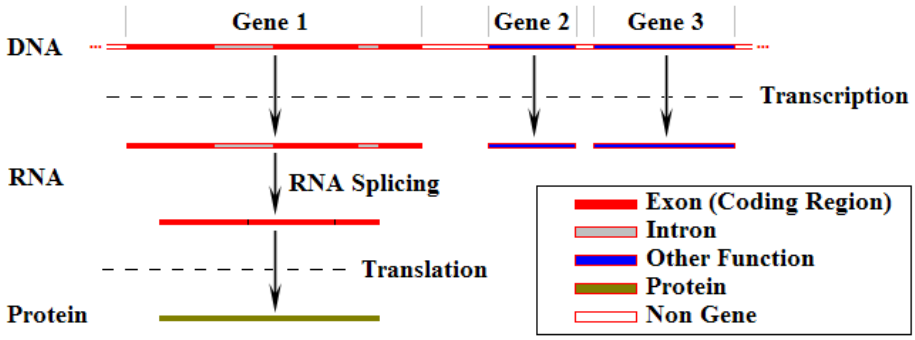


Figure 2.12 Protein synthesis

Table 2.1 The genetic code

1st position (5' end) ↓	2nd position				3rd position (3' end) ↓
	U (T)	C	A	G	
U (T)	Phe	Ser	Tyr	Cys	U (T)
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U (T)
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U (T)
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U (T)
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

In the central dogma, DNA replication is another important process, by which genetic information can be transmitted from parent DNA to progeny DNA. The replication of DNA helix begins with a local separation of its two strands. Each of the two strands serves as a template for the formation of a new complementary strand by the base pairing principle (A-T and G-C). Different from the transcription, the two new strands then remain as complementary strands of the original two strands, so that two new DNA helixes are generated (Alberts et al., 1994). Although transcription and translation are processes to translate DNA sequences into proteins, namely, the protein synthesis, DNA replication is also a critical process since it keeps the repetition of gene expression in generations of cells or organisms.

According to the central dogma, it is not the complete DNA sequence that directs the synthesis of proteins. Only some fragments (genes) are transcribed into RNA, in which only the mRNA (after splicing if any) are finally translated into proteins. Therefore, DNA fragments, which appear in the final mRNA and are finally translated into proteins, are called “protein-coding regions” or “coding regions” for short (Figure 2.12). It is also noticed that, in transcription, both the W strand and the C strand can be the template, so the base sequence in the mRNA can be the same as the W strand or the C strand, which is a complementary strand running in the opposite direction. Meanwhile, a certain region can be read in triplets beginning with the 1st base, the 2nd base, or the 3rd base. Considering these two factors (two reading

directions, three starting positions), the represented base sequence (the W strand sequence) of a coding region can be read in 6 possible ways, which are called 6 Opening Reading Frames (ORFs) (Cebrat and Dudek, 1996 and 1998; Anastassiou, 2001), as in the example shown in Figure 2.13. Hence, finding the correct ORF, in which the DNA sequence is read during protein synthesis, is another significant issue in predicting correct protein products.

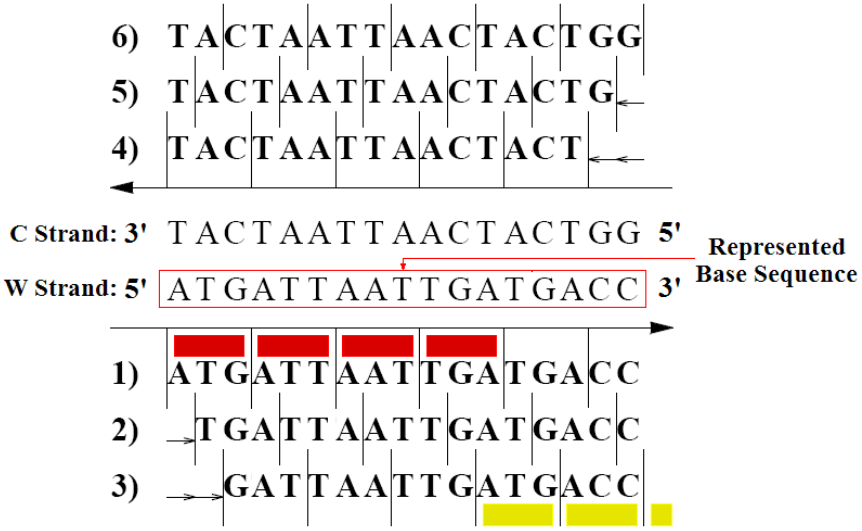


Figure 2.13 An example of the 6 Opening Reading Frames (ORFs)

2.3 Nucleosome and nucleosome positioning

Besides computational coding region prediction, another objective in this study is to investigate the relationship between sequence periodicity and nucleosome formation. Therefore, this section reviews the knowledge of nucleosome, as well as previous findings of nucleosome positioning.

2.3.1 Chromosome and nucleosome

DNA molecules are packaged into a smaller volume to fit into their cells. In the nuclear envelope of a eukaryotic cell, folded DNA molecules are contained in a structure named “chromosome”, which is a combination of DNA, histone, and some other proteins (Van Holde, 1989; Cooper, 2000). Besides packaging DNA molecules, the chromosome structure also strengthens DNA and controls gene expression and DNA replication (Cooper, 2000).

As shown in Figure 2.14, the basic structure of a eukaryotic chromosome has three levels (Cooper, 2000): DNA wraps around histone proteins to form nucleosomes, which is said to be the “beads on a string” structure. The second level of this structure is a 30nm condensed chromatin fiber consisting of nucleosome arrays in their most compact form. Finally, higher-level DNA is packaged into a metaphase chromosome.

As the most basic unit in a eukaryotic chromosome, a nucleosome consists of a histone core and a wrapped fragment of DNA helix (Luger et al., 1997; Segal et al., 2006; Nair, 2009; Jiang and Pugh, 2009). Each histone core is composed of two copies of each of the histone proteins H2A, H2B, H3, and H4 (Figure 2.15a). A ~147bp fragment of DNA helix is sharply bent and tightly coiled around the histone core for ~1.65 times in a left-handed toroid (Luger et al., 1997; Jiang and Pugh, 2009). As shown in Figure 2.15b, this sharp bending occurs at every DNA helical

repeat (~10bp) of this ~147bp fragment, when the major groove of the helix faces inwards, i.e., towards the histone core, and with an opposite direction at ~5bp away, when the major groove faces outward (Segal et al., 2006). Beyond nucleosomes are linkers, each consisting of a linker histone, i.e., H1, attached to a linker DNA fragment with a length of 10~50bp (Van Holde, 1989; Jiang and Pugh, 2009). They link nucleosomes so that nucleosomes are arranged in a linear array along the DNA polymer (Figure 2.15a). 75~90% of genomic DNA is wrapped in nucleosomes (Van Holde, 1989; Segal et al., 2006). Such basic structure has been found to be related to the regulation of gene expression. Access to DNA wrapped in a nucleosome is said to be occluded for polymerase, regulatory, repair, and recombination complexes (Richmond and Davey, 2003), but nucleosomes also recruit other proteins through interactions with their histone tail domains (Jenuwein and Allis, 2001). Meanwhile, it is reported in some literatures (Kamakaka and Biggins, 2005; Sarma and Reinberg, 2005; Jiang and Pugh, 2009) that the histone variants are related to the activation of genes: At active genes or at genes that are poised for activation, histones H2A and H3 are replaced by histone variants H2A.Z and H3.3 (Figure 2.15a). Therefore, the study on locations of nucleosomes along the DNA molecule and the principle of nucleosome formation are significant for learning about the regulation of gene expression (Kornberg and Lorch, 1999; Wyrick et al., 1999; Segal et al., 2006).

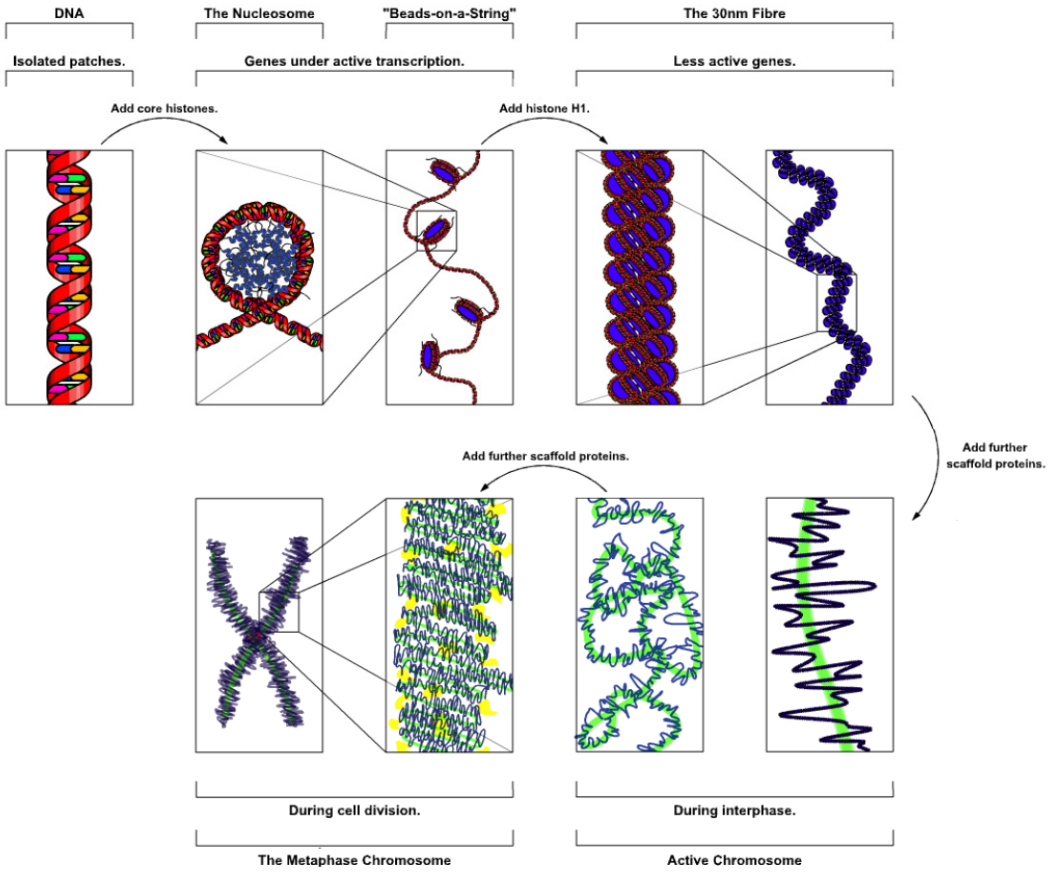


Figure 2.14 Basic structure of a eukaryotic chromosome

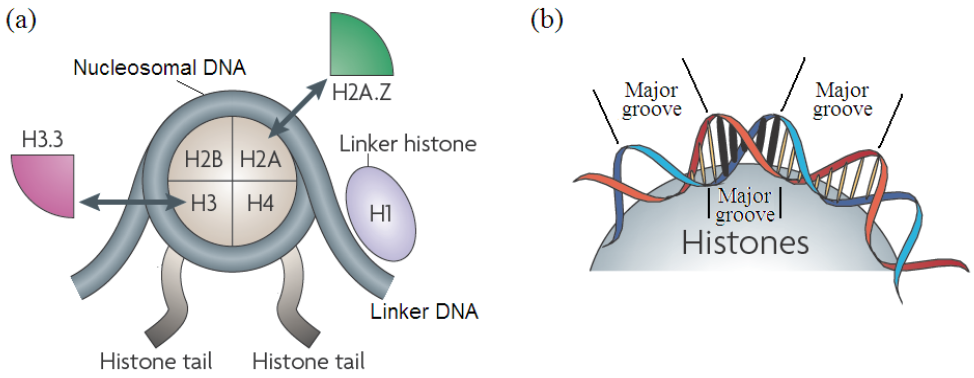


Figure 2.15 Structure of a nucleosome (Jiang and Pugh, 2009) (a) The structure consists of a histone core (H2A, H2B, H3, and H4) and a wrapped fragment of DNA helix (b) The DNA helix coils around the histone core

2.3.2 Nucleosome positioning

In the early years, locations of nucleosomes along DNA were studied in a low resolution (Jiang and Pugh, 2009). However, it was found there is a general depletion of nucleosomes in intergenic regions where promoters are found (Lee et al., 2004; Bernstein et al., 2004; Sekinger et al., 2005). Further studies confirmed that gene activation resulted in additional nucleosome depletion (Schwabish and Struhl, 2004; Guillemette et al., 2005; Zanton and Pugh, 2006). By 2005, higher resolution views of nucleosome locations were provided by the development of DNA microarrays (Jiang and Pugh, 2009). It is found that the nucleosomes at most genes are organized in the same way (Yuan et al., 2005): a Nucleosome-Free Region (NFR) is flanked by two well-positioned nucleosomes, i.e., the –1 and +1 nucleosomes, followed by a nucleosomal array that packages the gene (Figure 2.16). This basic pattern is also found in metazoans (Barski et al., 2007; Mavrich et al., 2008).

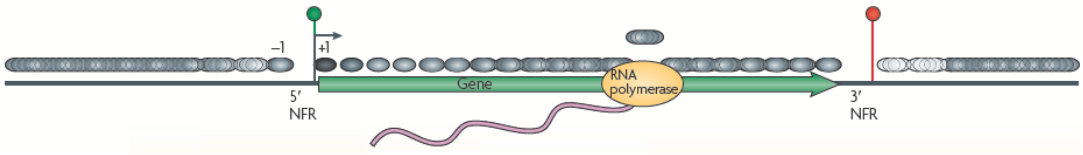


Figure 2.16 Organization of nucleosome locations at genes (Jiang and Pugh, 2009)

In 2007, a comprehensive map of nucleosome locations in a eukaryotic genome (*S. cerevisiae*) was completed, owing to two technological advances (Jiang and Pugh, 2009). First, the density of microarray probes increased dramatically, so that millions

of genomic loci can be interrogated by ChIP-chip analysis (Aparicio et al., 2004; Jiang and Pugh, 2009) in a single experiment. The distances between probes along DNA sequence were reduced to 5bp in *S. cerevisiae* and 36bp in *Drosophila melanogaster* (Lee et al., 2007; Mavrich et al., 2008). Second, individual nucleosomal DNA molecules were able to be sequenced, due to massively parallel shotgun sequencing.

By a ChIP-Seq method (Jiang and Pugh, 2009), high-resolution mapping of nucleosomes is provided (Figure 2.17). It involves an initial step to cross-link histones to nucleosomal fragments of experimental DNA molecules, by formaldehyde treatment of living cells. In principle, nucleosomes are trapped at their *in vivo* locations by this cross-linking step. After that, by digestion with high levels of Micrococcal Nuclease (MNase), linker DNA particles are removed, so that subsets of nucleosomal particles can be isolated by immunoprecipitation using antibodies directed against histones, histone variants, or histone modifications (like in the ChIP-chip). Nucleosomal fragments of DNA (~150bp) are further refined by a size-selection step using agarose gel electrophoresis (Berg et al., 2002). Then, the 5' ends of millions of individual DNA fragments in this library are sequenced in parallel. Short-read technology sequences 25~35bp fragments (called tags), while long-read technology produces read tags with lengths of 100bp or more. Sequence tags are then mapped to the reference DNA sequence, on either the Watson or Crick strand, using

alignment algorithms (Figure 2.17). The 5' ends of each tag, corresponding to nucleosome borders, are then plotted as a bar graph at each coordinate along the DNA sequence. Next, the nucleosome midpoints are represented by +73bp from the tag locations for the W strand and -73bp for the C strand (Albert et al., 2007; Mavrich et al., 2008). Clusters of tags show distribution of nucleosome positions in cells. The tighter the cluster, the more "phased" the corresponding nucleosome is. Randomly distributed tags reflect random ("fuzzy") positioning (see discussions on "phased" and "fuzzy" below).

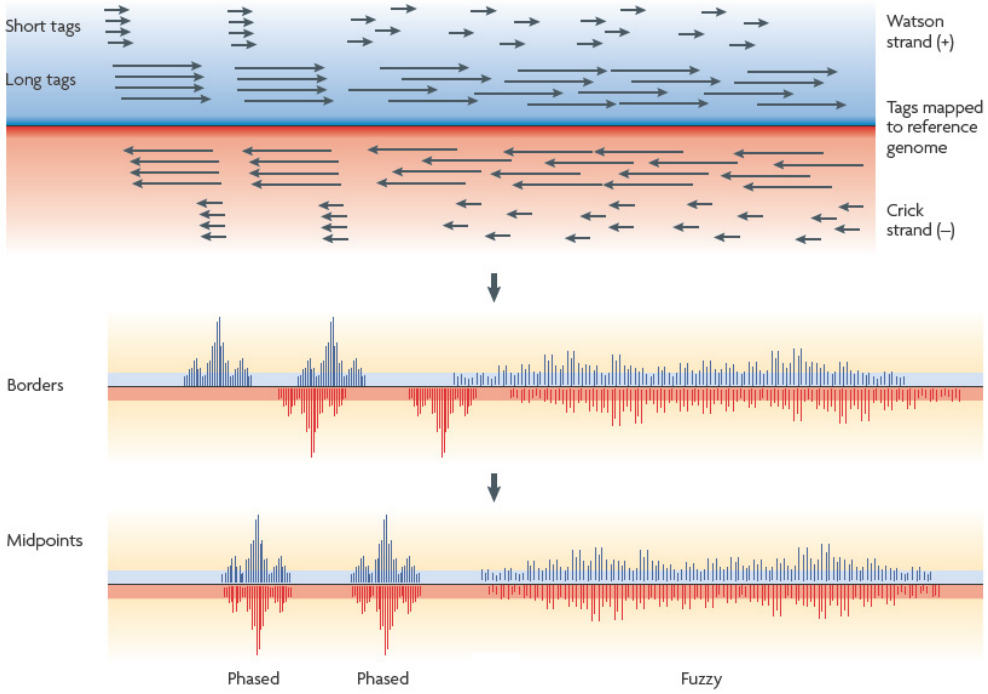


Figure 2.17 ChIP-Seq method for high-resolution nucleosomes mapping along a DNA sequence (Jiang and Pugh, 2009). The blue area stands for the Watson (W) strand, the red area stands for the Crick (C) strand, the short arrows indicate the short-read tags, the long arrows indicate the long-read tags, and the bar spectrums show the distribution of nucleosome positions in cells.

For the sequencing step in the above mapping process, there have been three DNA sequencing technologies used (Barski et al., 2007; Albert et al., 2007; Shivaswamy et al., 2008; Valouev et al., 2008; Mavrich et al., 2008; Jiang and Pugh, 2009): Pyrosequencing can sequence nucleosomal DNA end-to-end, so it is able to provide the greatest mapping accuracy, particularly in genomic regions of a low complexity. In contrast, other platforms, such as those provided by Illumina–Solexa Genome Analyzer and Applied Biosystems SOLiD, generate only 25~35bp sequence tags. However, these short-read technologies produce >100 times the number of sequence tags at a similar cost as the long-read technologies do, so they are currently the practical sequencing technologies for mapping nucleosomes in large genomes. The higher count of the short-read tags enhances mapping accuracy and thus provides a practical way of mapping nucleosomes.

By the genome-wide nucleosome mapping, the pattern of nucleosomes' distribution has been revealed. The locations, where nucleosomes are bound along a certain DNA sequence, may be different among different cells (Valouev et al., 2008). However, at any given genomic locus, the preference of nucleosome binding can be described (Figure 2.18): At most loci, there is an approximately normal distribution of nucleosome positions around particular genomic coordinate, ranging from ~30bp width for highly phased nucleosomes to a uniformly random distribution for fuzzy nucleosomes (Jiang and Pugh, 2009). Meanwhile, within each normal distribution

(highly phased locus), there are some "second-level" normal distributions ~10bp apart from each other (Figure 2.18), revealing a preference of nucleosome binding for a certain phase in the ~10bp period of the DNA helix (Albert et al., 2007).

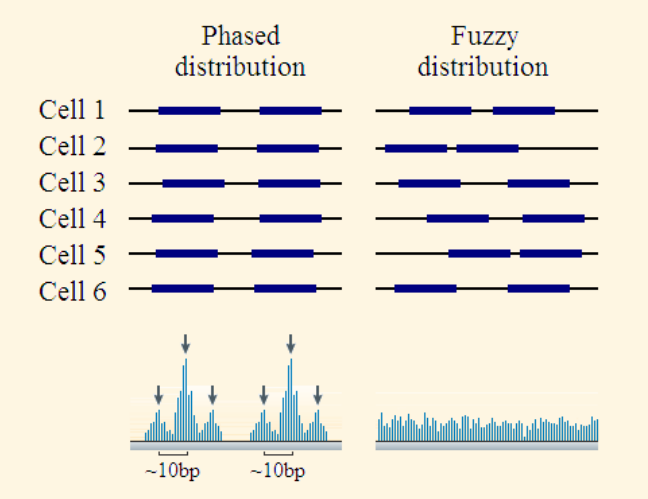


Figure 2.18 Distribution of nucleosome positions along a certain DNA sequence. The above is an example of the nucleosome locations in different cells. The blue bars indicate the positions of the nucleosomes. Below is the distribution spectrum of the nucleosome positions over all cells in the experiment. The arrows indicate the preferred phase positions in the ~10bp period at phased nucleosome loci.

2.4 Current methods for computational coding region prediction

Computational coding region prediction takes a DNA sequence as well as some other information (previous knowledge, experimentally detected protein data, and so on), if necessary, as its input, and provides a prediction of the locations of protein-coding regions along the DNA sequence. It is significant to review some current methods for this end, before developing a new approach in this study.

2.4.1 Investigation into coding related properties in DNA sequences

In the early years, there were already some investigations of the relationship between sequence properties and protein coding, such as Bennetzen and Hall's codon bias index (Bennetzen and Hall, 1982), Staden and McLachlan's measures based on differences in codon usage (Staden and McLachlan, 1982), and the hexamer counts (Claverie and Bougueleret, 1986). Most of them are concerned about the biased usage of codons.

And then, various meaningful statistic properties were extracted for the analysis. For instance, some entropy analyses on parts of genomes were conducted (Orlov et al., 2006), and the statistical dependencies between nucleotides were analyzed in many ways such as mutual information function (Li, 1997) and hidden Markov models (Stanke and Waack, 2003). In 2000, Zhang and Wang, based on the Z curve representation (Zhang and Zhang, 1994), proposed a measure called YZ score, to identify coding regions (Zhang and Wang, 2000).

Besides, Long-Range Correlation (LRC), also called Long-Range Dependence (LRD), is a significant coding related property concerned by many researchers. A time-series is long-range dependent if it has correlations which persist over all time scales (Clegg, 2006). In mathematics, LRD is defined as below. If $\{X_t \mid t = 1, 2, \ldots\}$ is a time-series which is weakly stationary (it has a finite mean and the autocovariance depends only on the separation between points in the series), the

Autocorrelation Function (ACF) of X_t is:

$$\rho(k) = \frac{E[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma^2}$$

Here, $E[X_t]$ is the expected value of X_t, μ is the mean value, and σ^2 is the variance. Then, the series is said to be long-range dependent if $\sum_{k=-\infty}^{+\infty} \rho(k)$ converges. And often, there is a specific functional form for such a $\rho(k)$, that is:

$$\rho(k) \sim C_\rho k^{-\alpha}$$

Here, $C_\rho > 0$ and $0 < \alpha < 1$. The symbol "~" means "asymptotically equal to", that is $f(x) \sim g(x) \leftrightarrow f(x) = g(x)$ as $x \to \infty$. It is said that constant α is related to a Hurst exponent (H) via the equation $\alpha = 2 - 2H$ (Clegg, 2006). Hence, H ranges from 0.5 to 1 in this case.

However, in the LRC analysis, the Hurst exponent is concerned more than the coefficient α, and a practical rescaled range statistics (R/S) method is proposed to estimate H instead of calculating α. As the main idea of the R/S method, H is estimated by investigating an R/S ratio (Devynck et al., 2000; Mandelbrot and Hudson, 2004). Consider any sub-sequence with a length of N from the time-series, which is $\{X_{t_0+k} \mid k = 1, 2, \ldots, N\}$. The R/S ratio of this sequence is defined as:

$$\frac{R(N)}{S(N)} = \frac{\max\{0, W_1, W_2, \ldots, W_N\} - \min\{0, W_1, W_2, \ldots, W_N\}}{\sqrt{S^2(N)}}$$

$$W_k = X_{t_0+1} + X_{t_0+2} + \ldots + X_{t_0+k} - k\overline{X}(N)$$

Here, $\overline{X}(N)$ and $S^2(N)$ are the sample mean and the sample variance of this sub-sequence respectively. Then a relationship is expected to exist among the R/S

ratio, the Hurst exponent H, and the data amount N, when $N \to \infty$:

$$\frac{R(N)}{S(N)} \sim cN^H \quad \text{or} \quad \log\left[\frac{R(N)}{S(N)}\right] \sim H \log N + \log c$$

Here c is a constant coefficient independent with the value of N. So, after calculating the R/S ratio over a number of sub-sequences with relatively high values of N, H can be estimated by a log-log regression on the (R/S, N) pairs. In this case, H is defined in a more general range (0, 1) rather than (0.5, 1). And it shows that there is a strong relationship between H and LRC: If $H = 0.5$, the series is random or with a local correlation (Peng et al., 1992); $H > 0.5$ or $H < 0.5$ indicates a persistent (to enhance the previous trend) or anti-persistent (to resist the previous trend) pattern of the series trend respectively (Devynck et al., 2000; Mandelbrot and Hudson, 2004).

In 1992, Peng et al. made research into LRC of DNA sequences. They conducted a Purine-Pyrimidine (PP) DNA walk. By considering the 1-dimensional walk as a time-series $\{Y_t \mid t = 1, 2, \ldots\}$, they calculated the Hurst exponent by an improved algorithm (Peng et al., 1992), which is based on the relationship between the variance of increments and the time separation (Feder, 1988):

$$F^2(\Delta t) = Var(Y_{t+\Delta t} - Y_t) \sim \Delta t^{2H}$$

Then Peng et al. found that PP walks of sequences in intron-containing genes are highly correlated ($H > 0.5$) while this correlation is absent in genes without any intron (Peng et al., 1992). This result was also confirmed by Haimovich et al. (2006). Moreover, in Haimovich et al.'s work, they further investigated the Hurst exponents

of random walks in the complex plane for DNA sequences containing tRNA and rRNA. And by their analysis on the Hurst exponents, they suggested that all the regions of biological interest (containing exons, introns, rRNA, rRNA, and so on) have different pattern structures from their surrounding segments (Haimovich et al., 2006).

Besides the Hurst exponent analysis on PP DNA walks, Te Boekhorst et al. (2008) also conducted a Weak-Strong (WS) DNA walk, in which the weak bond bases (A and T) and the strong bond bases (G and C) indicate "up" and "down" respectively. And then, based on the results of the LRC analysis on PP walks and WS walks, they constructed a 2-dimensional classifier to discriminate coding, non-coding, and regulatory regions (Te Boekhorst et al., 2008) as shown in Figure 2.19.

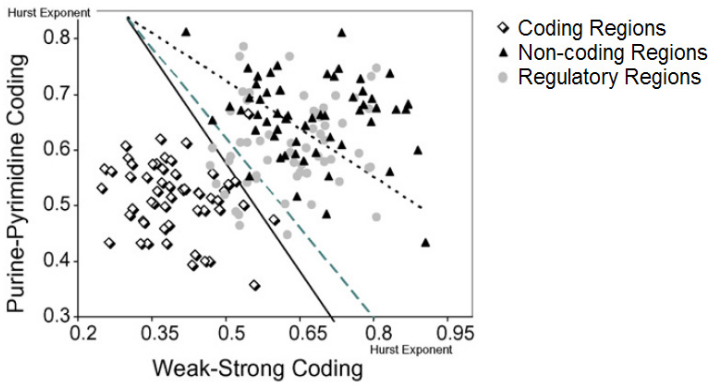


Figure 2.19 Te Boekhorst et al.'s 2-dimensional classifier (Te Boekhorst et al., 2008)

2.4.2 Popular programs for coding region prediction

Although many statistic properties related to protein-coding have been reported

in many literatures mentioned in the above section, such properties are too rough to complete the prediction with high accuracy, and few popular programs are directly based on these properties. In order to achieve high accuracy, most of them use detailed knowledge about sequence patterns learned from previously well-annotated sequences (the trained models) to find coding regions in new sequences. There is said to be two main classes of programs for computational coding region prediction. One is based on sequence similarity searching, while the other is based on structure recognition (Wang et al., 2004).

The sequence similarity searching based methods assume that functional regions, such as exons, are more conserved evolutionarily than nonfunctional regions (Wang et al., 2004), and meanwhile unconserved regions do not contribute to the similarity (Smith and Waterman, 1981; Goad and Kanehisa, 1982; Sellers, 1984). Such methods compare the query sequence with Expressed Sequence Tags (ESTs) or expressed proteins in a database, find regions with a high similarity with certain ESTs or proteins, and predict gene regions using the information of these findings. Here, ESTs stand for some short sub-sequences which have been found to be transcribed in previous experiments on well-studied organisms (Adams et al., 1991).

Local or global alignments are basic techniques for sequence similarity searching. The Basic Local Alignment Search Tool (BLAST) family of methods is most commonly used for fast local alignment among amino-acid sequences or

nucleotide sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences and identify library sequences that resemble parts of the query sequence. The first BLAST is proposed by Altschul et al. (1990). In this original BLAST, a similarity score of a local aligned segment is the sum of similarity values of aligned residue pairs in the segment, and the similarity value for any nucleotide (or amino acid) pair is given by a score matrix, e.g. the PAM-120 matrix (Dayhoff et al., 1978) for amino acid sequences. The BLAST complete sequence similarity searching in 3 steps: First, a word list is built by finding all possible words with a length of ω and a similarity score no less than T when compared with words in the query sequence. Second, the BLAST searches the database for sequences containing the words (the hits). Finally, each hit is extended to find the Maximal Segment Pair (MSP) of the database sequence and the query sequence. To extend a hit, the BLAST terminates the process of extending in one direction when a segment pair is reached whose score falls a certain distance below the best score found from shorter extensions. According to Altschul et al. (1990), Dynamic Programming (DP) can also be used to extend hits so as to allow gaps in the resulting alignments. But, this greatly slows the extension process and reduces the selectivity.

Besides, some programs conduct sequence similarity searching based on global alignment. For instance, PROCRUSTES (Gelfand et al., 1996) and GeneWise

(Birney and Durbin, 2000), use global alignment between homologous proteins and translated ORFs in genomic sequences for gene prediction. The biggest limitation to methods based on sequence similarity searching is that only about half of the genes being discovered have significant homology to genes in databases. So some more flexible programs of this type also combine structure recognition methods, for instance, GeneParser (Snyder and Stormo, 1993; Snyder and Stormo, 1995), Genie (Kulp et al., 1996; Reese et al., 1997), and GRAIL (Uberbacher and Mural, 1991).

The structure recognition based methods use the structural features of known genes as templates to detect new genes. They relay on two types of sequence information: signal information and content information (Burge and Karlin, 1998; Stormo, 2000; Wang et al., 2004; Do and Choi, 2006). Signal information stands for short motifs in sequences, including start codons, stop codons, splice sites, branch points, polypyrimidine tracts, and so on. Such information helps with the detection of exact boundaries of different sequence regions. Content information refers to some statistical properties in longer local regions, such as the patterns of codon usage which are unique to a certain species. This allows coding sequences to be roughly distinguished from the surrounding non-coding sequences in statistics.

Programs with the idea of structure recognition have been developed using various algorithms, including Dynamic Programming (GeneID: Guigó et al., 1992. GRAIL: Uberbacher and Mural, 1991), discriminant analysis (MZEF: Zhang, 1997),

Linguist methods (GenLang: Dong and Searls, 1994), Hidden Markov Models (Borodovsky and McIninch, 1993; Burge and Karlin, 1997; Salzberg et al., 1998; Guigó et al., 2000; Do and Choi, 2006), and Neural Networks (GeneParser: Snyder and Stormo, 1993; Snyder and Stormo, 1995. GRAIL: Uberbacher and Mural, 1991). Among them, the Hidden Markov Model (HMM) based programs are said to be the most successful. An HMM is a statistical Markov model in which the system is assumed to be a Markov process with unobserved (hidden) states and a series of observations (Rabiner, 1989; Huang et al., 1990; Ephraim and Merhav, 2002). An HMM can be considered as the simplest dynamic Bayesian network, describing outputs from states to observations and transitions between hidden states. As in Figure 2.20a, in an HMM, at any step t in the process, state s may transfer to another state s' following a probability of $a_{ss'}$, and meanwhile state s may output an observation o following a probability of b_{so}. It must be noticed that all the states in HMM are unobservable (hidden), which can only be estimated by the previously established (trained) model and the series of observations. As a simplest example of HMM for gene prediction mentioned in review paper (Wang et al., 2004), the states are designed as "Up stream", "Start Codon", "Exon", "Stop Codon", "Down stream", "3' Splice site", "5' Splice site", and "Intron", and the topologic relationship between them is described in Figure 2.20b. Suppose a DNA sequence (the observations series) with a length of L is given as $O = \{o_1, o_2, \ldots, o_L\}$. The hidden state series is $S = \{s_1,$

$s_2, \ldots, s_L\}$. Then the conditional probability of S, given that the observation output is O, can be computed using Bayes' Rule as:

$$P\{S|O\} = \frac{P(S,O)}{\sum_{S' \in \Phi_L} P(S',O)} = \frac{\lambda_{s_1} \cdot a_{s_1 s_2} \cdot a_{s_2 s_3} \cdot \ldots \cdot a_{s_{L-1} s_L} \cdot \varphi_{s_L} \cdot b_{s_1 o_1} \cdot b_{s_1 o_1} \cdot \ldots b_{s_L o_L}}{\sum_{S' \in \Phi_L} P(S',O)}$$

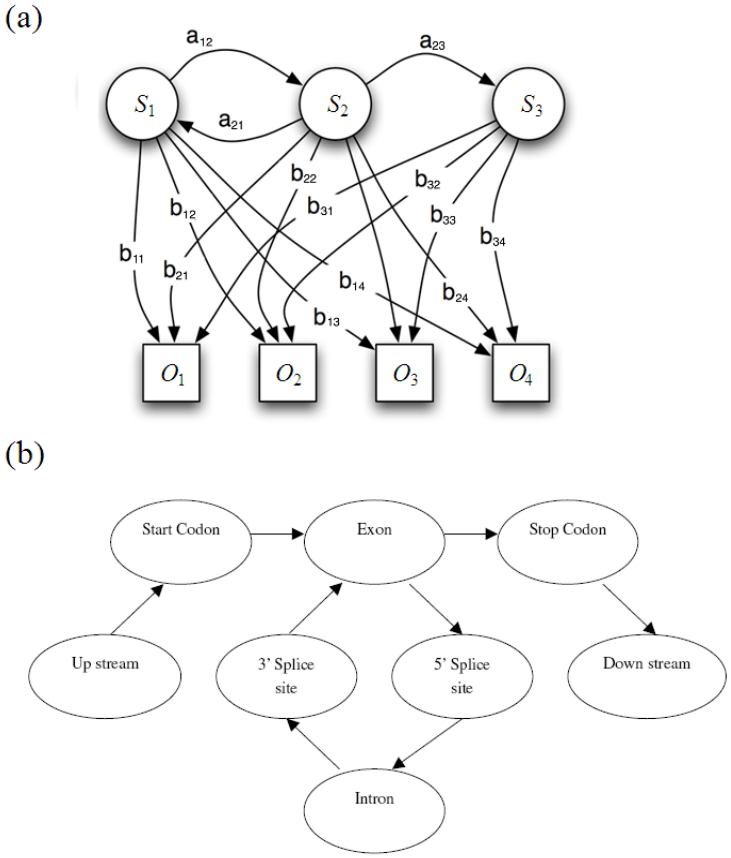


Figure 2.20 Hidden Markov Model (HMM) (a) The Bayes network of an HMM (b) The topologic relationship between states in the simplest HMM for gene prediction (Wang et al., 2004)

Here, Φ_L is the set of all possible state series S' with a length of L, the parameter λ_s is the probability that a state series begins with the state s, ψ_s is the

probability that a state series ends with the state s, $a_{ss'}$ is the probability of transition from state s to state s', and b_{so} is the probability of outputting observation o from state s. All these parameters need to be estimated by a preceding training process. Then, given a particular DNA sequence O, a state series S can be found that maximizes the likelihood of generating O. In other words, for a particular sequence, we can find the functional units that are most likely to represent the sequence. Therefore, the model can be used for automatic annotation of DNA sequences (Wang et al., 2004).

From the original HMM, many extensions have been proposed specifically for gene prediction, such as Generalized Hidden Markov Model (GHMM) (Kulp et al., 1996) and class HMM (CHMM). Based on HMM, gene predicting programs are developed, including FGENESH (Solovyev et al., 1994), Genie (Kulp et al., 1996; Reese et al., 1997), GENSCAN (Burge and Karlin, 1997; Salzberg et al., 1998), HMMgene (Krogh, 1997), and GeneMark.hmm (Borodovsky and McIninch, 1993). Prediction accuracies of such methods are extremely high, evaluated by *ab initio* experiments (Guigó et al., 2000). However, the critical dependence on the training process may reduce adaptability of these methods, particularly, for new sequences from unknown organisms with no or small training sets (Do and Choi, 2006). Therefore, it is significant to develop a "training-free" method to eliminate this defect.

2.4.3 Methods based on the TP property

The phenomenon of the Triplet Periodicity (TP) in DNA sequences, as mentioned in Section 1.1, was first connected with genetic coding by Fickett in 1982. In 1992, Fickett and Tung proposed a position asymmetry measure to extract the TP property (Fickett and Tung, 1992). Denote a DNA sequence with a length of N as $\{X_t \mid t = 1, 2, \ldots, N\}$, and for each position t, a so-called test-codon position number is $r = t \bmod 3$. A function $f(\Lambda, r)$ is defined to record the frequency by which base Λ (A, T, G, or C) appears in each test-codon position r. And then, for each base Λ, the position asymmetry measure, i.e., asymm(Λ), is defined as:

$$\text{asymm}(\Lambda) = \sum_{r=0}^{2} [f(\Lambda, r) - \mu(\Lambda)]^2 \quad (\Lambda = A, T, G \text{ or } C)$$

$$\mu(\Lambda) = \frac{1}{3} \sum_{r=0}^{2} f(\Lambda, r)$$

It is noticed that asymm(Λ) is actually the variance of $f(\Lambda, r)$ over three test-codon positions. A high value of asymm(Λ) indicates a highly biased appearance of Λ, which suggests a high intensity of TP.

Tiwari et al. (1997) developed a measure known as Spectral Content Measure to construct a gene predictor. In this method, a DNA sequence is represented as four binary sequences of Voss's model (Voss, 1992), i.e., $u_A(t)$, $u_T(t)$, $u_G(t)$, and $u_C(t)$. Consider the Fourier Transform on the four sequences, which is:

$$U_\Lambda(k) = \sum_{t=1}^{N} u_\Lambda(t) e^{-i\frac{2\pi}{N}tk}$$

Here, N stands for the length of the sequence. Four complex numbers can be

obtained at frequency $k = N/3$, i.e., $U_A(N/3)$, $U_T(N/3)$, $U_G(N/3)$, and $U_C(N/3)$. The Spectral Content Measure (SCM) is the square sum of these four components:

$$SCM = S(\frac{N}{3}) = \left|U_A(\frac{N}{3})\right|^2 + \left|U_T(\frac{N}{3})\right|^2 + \left|U_G(\frac{N}{3})\right|^2 + \left|U_C(\frac{N}{3})\right|^2$$

The measure is said to be the same as the sum of the four position asymmetry measures (up to a 3/2 multiplicative factor) (Kotlar and Lavner, 2003). So a high value of SCM also suggests a high intensity of TP.

In 2000, Anastassiou introduced an Optimized Spectral Content Measure (OSCM) by assigning four optimized weights to the four complex components (Anastassiou, 2000 and 2001):

$$OSCM = \left|aU_A(\frac{N}{3}) + tU_T(\frac{N}{3}) + gU_G(\frac{N}{3}) + cU_C(\frac{N}{3})\right|^2$$

Here, weights a, t, g, and c are obtained by training using optimization technique applied to the known gene dataset. Rather than simply considering only the intensity of the TP property, by means of involving the four weights, OSCM considers a TP profile of the target organism, which can be considered as a more specific content information in the sequence. And in practice, OSCM is said to be more significant than SCM for predicting genes in *S. cerevisiae* (Anastassiou, 2000).

Kotlar and Lavner (2003) further proposed a Spectral Rotation Measure (SRM) by considering the phase angles of the four complex components, i.e., $\arg[U_A(N/3)]$, $\arg[U_T(N/3)]$, $\arg[U_G(N/3)]$, and $\arg[U_C(N/3)]$. They found that for a given organism,

the distribution of phase angles of each component, i.e., arg[$U_\Lambda(N/3)$], is bell-shaped for coding regions, and close to uniform for non-coding regions (Figure 2.21). This finding suggests that the pattern of the base appearance in triplets (codons) is highly persistent in coding regions, compared with the randomness in non-coding regions. Hence, different from non-coding regions, for a given coding region, the four components can be rotated to a same direction (the positive real direction) by four multiplications:

$$U_\Lambda(\frac{N}{3}) \to e^{-i\mu_\Lambda} U_\Lambda(\frac{N}{3}) \quad (\Lambda = \text{A, T, G, or C})$$

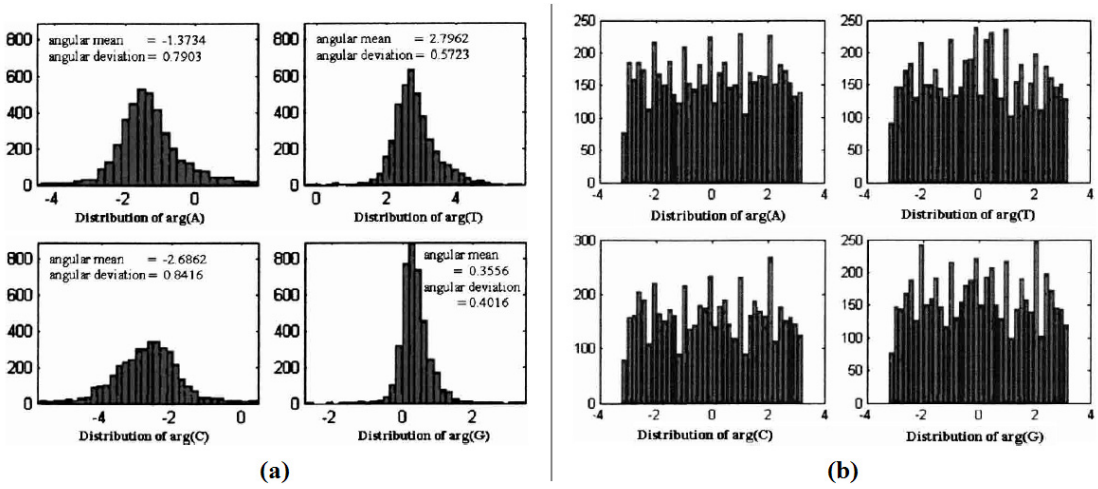


Figure 2.21 Distributions of phase angles of the four components (Kotlar and Lavner, 2003) (a) For coding regions (b) For non-coding regions

Here, μ_Λ is the expected value of arg[$U_\Lambda(N/3)$] obtained by the calculation from the known gene dataset. Then the Spectral Rotation Measure (SRM) is defined by the rotated components:

$$SRM = \left| \frac{e^{-i\mu_A}}{\sigma_A} U_A(\frac{N}{3}) + \frac{e^{-i\mu_T}}{\sigma_T} U_T(\frac{N}{3}) + \frac{e^{-i\mu_G}}{\sigma_G} U_G(\frac{N}{3}) + \frac{e^{-i\mu_C}}{\sigma_C} U_C(\frac{N}{3}) \right|^2$$

Here, σ_Λ stands for the variance of $\arg[U_\Lambda(N/3)]$ obtained by the calculation from the known gene dataset, and it is used in the measure to give more weight to narrower distributions (Kotlar and Lavner, 2003). A high value of SRM reveals a coding region, since only in coding regions can the four components be rotated to a same direction and produce a high summation. Kotlar and Lavner suggested that considering the arguments (phase angles) of the Fourier spectra yields more information than the corresponding magnitudes alone (Kotlar and Lavner, 2003).

And then a similar measure about the phase angles was adopted by Masoom et al. (2006) for detecting frame shifts in DNA sequences. Tuqan and Rushdi (2006) improved the SRM and proposed a Filtered Spectral Rotation Measure by bringing in a filter during the calculation. Jiang et al. (2008) developed a measure similar to the SCM by using a complex number sequence instead of the four binary sequences in Tiwari's method. Spectral Content related measures mentioned above have grown into a big family as researchers extended and improved Tiwari's original Spectral Content Measure in many ways (Yin and Yau, 2007; Tuqan and Rushdi, 2008; Chang et al., 2008; Akhtar et al., 2008a and 2008b; Ré and Pavesi, 2009).

For searching (detecting) coding regions in a complete DNA sequence by Spectral Content related measures, a Short Time Fourier Transform (STFT) is frequently used in practice, which means a Fourier Transform on local sections of the

complete sequence within a slide window. Fuentes et al. (2006) developed a significant recursive algorithm in order to reduce the computational load of the STFT with a slide window moving over the complete DNA sequence. Consider a number sequence $\{x_t \mid t = 1, 2, \ldots, N\}$ and a moving slide window on it with a length of L. When the window is at position t_0, the local section in the slide window is $\{x_{t_0}, x_{t_0+1}, \ldots, x_{t_0+L-1}\}$. Then the STFT is:

$$X_{t_0}(k) = \sum_{n=0}^{L-1} x_{t_0+n} e^{-i\frac{2\pi}{L}nk}$$

Then, in the next moving step, namely, when the window is at position t_0+1, we have:

$$X_{t_0+1}(k) = \sum_{n=0}^{L-1} x_{t_0+1+n} e^{-i\frac{2\pi}{L}nk} = e^{i\frac{2\pi}{L}k} \sum_{m=1}^{L} x_{t_0+m} e^{-i\frac{2\pi}{L}mk} \qquad (m = n+1)$$

$$= e^{i\frac{2\pi}{L}k} \left(x_{t_0+L} - x_{t_0} + \sum_{m=0}^{L-1} x_{t_0+m} e^{-i\frac{2\pi}{L}mk} \right) = e^{i\frac{2\pi}{L}k} \left[x_{t_0+L} - x_{t_0} + X_{t_0}(k) \right]$$

Hence, a recursive relationship is found between the two moving steps, i.e., $X_{t_0}(k)$ and $X_{t_0+1}(k)$. Based on this principle, by recursively calculating $X_{t_0+1}(k)$ from $X_{t_0}(k)$, Fuentes et al. developed their algorithm and reduced the computational complexity to $O(L+N)$, where N is the length of the complete sequence (Fuentes et al., 2006).

Besides Spectral Content related measures mentioned above, there are still many other methods using the TP property to complete coding region prediction. In 1996, different from using the four binary sequences, Dodin et al. generated another

digital sequence $S(k)$ from a DNA sequence by their algorithm (Dodin et al., 1996). For a base sequence $\{x_t \mid t = 1, 2, \ldots, N\}$, its corresponding $S(k)$ sequence is generated as:

$$S(k) = \frac{\sum_{i=1}^{N-1-k} g(i, i+1+k)}{N-1-k} \qquad (k = 0, 1, 2, ..., N-2)$$

where

$$g(i,j) = \begin{cases} 1 & x_i = x_j \\ 0 & x_i \neq x_j \end{cases}$$

When Dodin et al. conducted Fourier Transform on $S(k)$, they also observed the TP property. Thus, they claimed that Fourier Transform and Wavelet Transform on sequence $S(k)$ could be a tool to help visualize regular patterns in DNA sequences (Dodin et al., 2000).

Yan et al. (1998) conducted Fourier Transform on the Z curve and investigated the TP property. Cao et al. (2005) proposed a codon index revealing the TP by considering the recurrence time statistics of sequences. Also in 2005, Gao et al. claimed that, during the analysis of LRC, extra deviation of the fitting curve might be induced by the TP property, and they investigated the TP indirectly by observing this extra deviation (Gao et al., 2005). Korotkov et al. (2003) used an $n \times k$ matrix to extract the n-symbol periodicity in a sequence with k-size alphabet. By considering the mutual information, they derived a Z value as a measure to detect hidden periodicities in symbol sequences. Based on Korotkov's matrix, Frenkel and

Korotkov (2008) proposed a 3×4 Triplet Periodicity Matrix (TPM) to extract the information of the TP profile in DNA sequences. They conducted coding region identification by estimating the similarity between TPMs of the concerned DNA fragment and an artificial sequence with a complete triplet periodicity. In 2009, they further investigated frame shifts by using their TPM (Frenkel and Korotkov, 2009).

Most of the currently used methods based on the TP property employ certain measures to discriminate coding and non-coding regions. Such "measure-based" methods perform excellently in determining whether a sub-sequence is a coding region or not, and they are also applied to search coding regions in a complete sequence. For the searching purpose, a moving slide window is used and the measure of the local section in the slide window at every position is calculated. After that, a threshold is set to filter out coding regions. Such a scheme is commonly used, since it is available as long as a certain measure can be calculated. However, its defect for the searching purpose is obvious. In DNA sequences, coding regions can be with various lengths and coding-related properties can be in various scales. But in this scheme, the length of the slide window is fixed. It means that only the measure in a fixed scale is concerned. An unsuitable length of the slide window may reduce the sensitivity in detecting coding-related properties and lead to failure. Therefore, a flexible method is demanded, which can help easily detect coding-related properties in various scales.

2.5 Current arguments on nucleosome formation

This section reviews debated issues on nucleosome formation, as the background of the study on the relationship between sequence periodicity and nucleosome formation.

For long, nucleosome formation was thought to occur randomly along DNA molecules (Nair, 2009). Nowadays, owing to the development of the high-resolution nucleosome mapping, special patterns of nucleosome positioning have been found as mentioned in the above section. Researchers turn to study the principle of nucleosome formation, especially for highly phased nucleosomes.

Some researchers believe that the formation of nucleosomes is regulated by the cell and not significantly related to the DNA sequence (Ghaemmaghami et al., 2003; Cairns, 2005). Their explanation is that nucleosome positions might be regulated in the cell by abundant ATP-dependent nucleosome remodeling complexes, which move nucleosomes to locations whenever needed. But, on the other hand, special DNA sequence patterns are actually found in nucleosomal DNA fragments, suggesting that the formation of nucleosomes might be "sequence-specific" (Satchwell et al., 1986; Widom, 2001; Kiyama and Trifonov, 2002; Segal et al., 2006; Kaplan et al., 2009; Segal and Widom, 2009; Trifonov, 2010; Chen et al., 2010). The intrinsic DNA curvature raised by signals in DNA sequences has mostly attracted researchers' attention. Experimentally, a DNA curvature can be detected during polyacrylamide gel electrophoresis (Marini et al., 1982; Nair et al., 1994). Mobility

of a DNA molecule in the gel is said to be directly related to the mean square of its end to end distance (De Gennes, 1999). An elegant gel electrophoretic permutation assay has been designed to find the location of intrinsically curved DNA fragment (Wu and Crothers, 1984). A theoretical model for DNA curvature by De Santis et al. shows that curvature dispersion is linearly correlated with gel electrophoretic retardation (De Santis et al., 1988; De Santis et al., 1990; Zuccheri et al., 2001). There are mainly two models explaining how a DNA sequence instructs the curvature of the DNA helix and the nucleosome formation (Nair, 2009): One is as Trifonov's explanation (Trifonov and Sussman, 1980; Trifonov, 1980; Ulanovsky and Trifonov, 1987), in which the periodical appearance of hypothetical "wedges" along the DNA sequence curves the DNA helix. The other, which is called "junction bending model", attributes DNA curvature to the distortions at the junction between different DNA structural forms (Marini et al., 1982; Levene et al., 1986; Crothers et al., 1992). However, both models agree that the overall curvature is additive over the individual bending elements (Nair, 2009).

Many researches have been conducted to investigate significant bending elements in DNA sequences. As the most significant nucleosome-related sequence pattern, the ~10bp periodicity of dinucleotides was first discovered by Trifonov (Trifonov and Sussman, 1980). According to Trifonov's explanation (Trifonov and Sussman, 1980; Trifonov, 1980; Ulanovsky and Trifonov, 1987), the DNA helix is

bended because some dinucleotides, consisting of two nonparallel adjacent base pairs, serve as “wedges” (Figure 1.2b) regularly inserted into the DNA molecule, with a period of ~10bp. Therefore, a ~10bp periodicity of dinucleotides can be observed in the nucleosomal DNA fragments. After Trifonov’s work, many researches reported the periodicity property in DNA sequences related to nucleosome formation (Davey et al., 1995; Hosid et al., 2004; Albert et al., 2007; Mavrich et al., 2008; Chen et al., 2008; Salih et al., 2008; Nair, 2009; Jiang and Pugh, 2009). By investigation into the nucleosomal DNA fragments (Figure 2.22), the ~10bp periodicity of dinucleotides AA, TT, TA, and GC is most frequently reported to be related to nucleosome formation (Segal et al., 2006; Chen et al., 2010; Takasuka and Stein, 2010). The occurrence of AA and TT is commonly known to intrinsically curve the DNA axis, while $(CA)_n$ or $(CG)_n$ form Z-DNA structures (Kiyama and Trifonov, 2002; Nair, 2009). More generally, as shown in Figure 2.23, in nucleosomal DNA fragments, dinucleotides WW (W = A or T) are found to prefer to be placed where the minor groove is facing outward, while dinucleotides SS (S = C or G) prefer where the minor groove is facing inward (Trifonov, 2010). Therefore the ~10bp (the distance between two inward/outward minor grooves) periodicity of such dinucleotides is said to be related to nucleosome formation. However, there are still many different debates. For instance, Takasuka and Stein (2010) investigated the curvature of some synthetic DNA fragments with pervious reported motifs related to nucleosome

formation. Surprisingly, only those with the ~10bp periodicity of dinucleotide TA were found with high bendability and nucleosome-forming abilities. This result challenges other previous findings (Takasuka and Stein, 2010).

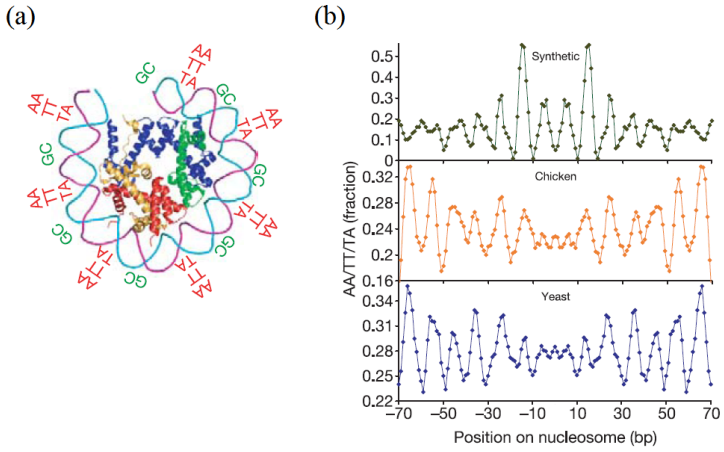


Figure 2.22 Periodical bending elements in a DNA helix (a) Dinucleotides AA, TT, TA, and GC curve the DNA helix (Segal et al., 2006) (b) The appearance fraction of dinucleotides AA, TT, and TA at each position of nucleosomal DNA fragments for yeast, chicken (Satchwell et al., 1986; Segal et al., 2006), or random chemically synthesized nucleosome-bound DNA sequences (Lowary and Widom, 1998; Segal et al., 2006). All nucleosomal fragments are aligned to the center (coordinate 0). It shows the ~10bp periodicity of these dinucleotides.

Meanwhile, some longer patterns are also found to be related to nucleosome formation (Figure 2.23). Consensus repeating patterns have been found in nucleosomal DNA fragments, such as $(YYYYYRRRRR)_n$ or particularly $A(TTTCCGGAAA)_nT$ (Johnson et al., 2006; Salih et al., 2008; Trifonov, 2010). Here, R = G or A, and Y = C or T. The sequence CC(GGRAATTYCC)GG is considered to be a theoretically predicted common pattern of DNA bendability in nucleosomes

(Nair, 2009; Trifonov, 2010). However, researchers argue that occurrence of this entire motif in eukaryotic sequences would result in formation of a very strong nucleosome. Therefore, it is practically avoided, e.g., in human genome, the motif GGAAATTTCC occurs only once or twice per million sites. Reduced versions of this motif are actually found. For example, trinucleotide components of patterns, GAA, AAA, AAT, ATT, TTT, and TTC, are found to be the most frequent triplets in all major eukaryotic genomes (Costantini and Bernardi, 2008; Trifonov, 2010).

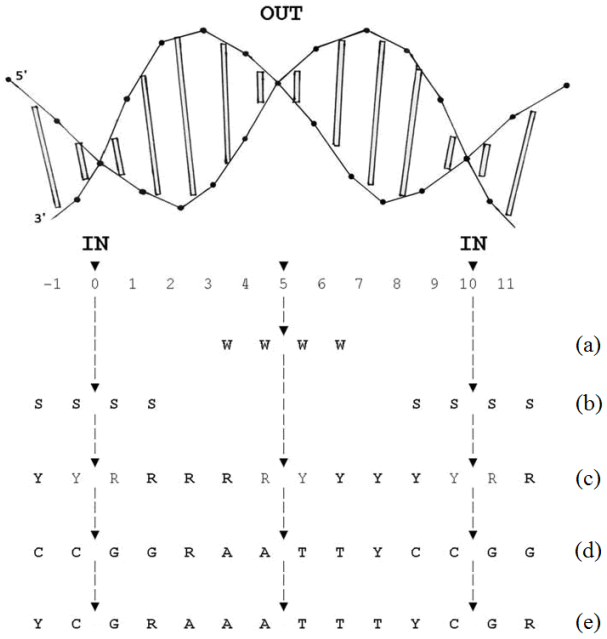


Figure 2.23 Positions of some special sequence patterns in nucleosomal DNA fragments (Trifonov, 2010) (a) dinucleotides WW (W = A or T) are placed where the minor groove is facing outward (b) dinucleotides SS (S = C or G) are placed where the minor groove is facing inward (c) pattern $(YYYYYRRRRR)_n$ curves the DNA helix (Y = C or T; R = A or G) (d) a theoretically predicted common pattern of DNA bendability in nucleosomes, i.e., $(GGRAATTYCC)_n$ (e) practical DNA bendability pattern $(GRAAATTTYC)_n$ in *C. elegans* nucleosomes

Although many results have been obtained by previous researches, the principle of nucleosome formation remains unclear and debates (such as whether the nucleosomes are formed under the regulation of the cell or not, how the periodicity in DNA sequences is related to nucleosome formation, and so on) still continue. Therefore, it is meaningful to conduct further studies on the relationship between sequence patterns and nucleosome formation.

2.6 Summary

In this chapter, comprehensive literature reviews on related issues in molecular biology have been given for a better understanding of the rest parts of this thesis. It began with a brief introduction to DNA, including DNA structure, DNA sequence, and various presentations of DNA sequence. After that, some necessary knowledge about gene and gene expression, mainly the "Central Dogma", was presented. Then reviews were given to the knowledge of nucleosome as well as previous findings on nucleosome positioning.

Besides the above background knowledge of molecular biology, current methods for computational coding region prediction have been reviewed, including TP-based methods (such as SCM, OSCM, SRM, and so on) and non-TP methods (similarity searching based or structure recognition based). This review reveals that although such methods are well-developed, most of them are critically dependent on

the training process or the fixed analysis scale. A flexible and “training-free” method is demanded to eliminate such defects.

In the end of this chapter, the main arguments on nucleosome formation have been presented, as the background of the study on the relationship between sequence periodicity and nucleosome formation. It shows that although many results have been obtained by previous researches, the principle of nucleosome formation remains unclear and debates still continue. Therefore, it is meaningful to conduct further studies on the relationship between sequence patterns and nucleosome formation.

CHAPTER 3

A NEW SELF-ADAPTIVE SPECTRAL ROTATION APPROACH FOR CODING REGION PREDICTION

As mentioned in Chapter 2, current methods for computational coding region prediction have disadvantages, such as training dependence and inflexible analysis scale. To eliminate such defects and improve the prediction, a new approach will be proposed in this chapter.

3.1 Self-Adaptive Spectral Rotation (SASR)

This section introduces a new approach named Self-Adaptive Spectral Rotation (SASR) to visualize the TP property in DNA sequences. The mathematical basis and the algorithm of this new approach are presented in detail.

3.1.1 TP vector

In Tiwari's method (Tiwari et al., 1997), a DNA sequence is represented as four binary strings, i.e., $u_A(t)$, $u_T(t)$, $u_G(t)$, and $u_C(t)$, by Voss's representation (Voss, 1992). $u_\Lambda(t) = 1$ if and only if base Λ (A, T, G, or C) appears at position t in the DNA sequence ($t = 1, 2, \ldots, N$). The Fourier Transform on these binary sequences gives:

$$U_\Lambda(k) = \sum_{t=1}^{N} u_\Lambda(t) e^{-i\frac{2\pi}{N}tk}$$

The values at the frequency $k = N/3$ are concerned:

$$U_\Lambda(\frac{N}{3}) = \sum_{t=1}^{N} u_\Lambda(t) e^{-i\frac{2\pi}{3}}$$
$$= \sum_{t \bmod 3=0} u_\Lambda(t) + e^{-i\frac{2\pi}{3}} \sum_{t \bmod 3=1} u_\Lambda(t) + e^{-i\frac{4\pi}{3}} \sum_{t \bmod 3=2} u_\Lambda(t)$$

So $U_\Lambda(N/3)$ can be viewed as a weighted summation of three complex numbers:

$$U_\Lambda(\frac{N}{3}) = F_{\Lambda 0} + F_{\Lambda 1} e^{-i\frac{2\pi}{3}} + F_{\Lambda 2} e^{-i\frac{4\pi}{3}} \tag{3-1}$$

where the weights:

$$F_{\Lambda j} = \sum_{t \bmod 3=j} u_\Lambda(t) \qquad j = 0, 1, 2$$

Therefore a triplet $\{F_{\Lambda 0}, F_{\Lambda 1}, F_{\Lambda 2}\}$ can be used as an equivalence to the Fourier Spectrum $U_\Lambda(N/3)$.

Meanwhile, in Frenkel and Korotkov's work (Frenkel and Korotkov, 2008; Frenkel and Korotkov, 2009), the TP profile was presented using a Triplet Periodicity Matrix (TPM). TPM is a 4×3 matrix, each row i stands for a nucleotide base (A, T, C, or G), each column stands for a position j (j = 1, 2, 3) in the 3bp period, and the entry m_{ij} is the count by which the base i appears at the position j. That is, for each row i = Λ:

$$\begin{aligned} m_{\Lambda 1} &= \sum_{t \bmod 3=1} u_\Lambda(t) = F_{\Lambda 1} \\ m_{\Lambda 2} &= \sum_{t \bmod 3=2} u_\Lambda(t) = F_{\Lambda 2} \\ m_{\Lambda 3} &= \sum_{t \bmod 3=0} u_\Lambda(t) = F_{\Lambda 0} \end{aligned} \tag{3-2}$$

Therefore, for each nucleotide base Λ, the triplet row vector $M = \{m_{\Lambda 1}, m_{\Lambda 2},$

$m_{\Lambda 3}\} = \{F_{\Lambda 1}, F_{\Lambda 2}, F_{\Lambda 0}\}$ from Frenkel and Korotkov's TPM is an equivalence to the Fourier spectrum $U_{\Lambda}(N/3)$. In this study, this triplet row vector M is called a TP vector and its corresponding Fourier spectrum is deemed as its complex form. The TP vector of a given DNA sequence X for nucleotide base Λ is then denoted here in a function form: $M_{\Lambda}(X)$. An example in Figure 3.1 shows the calculation of TP vectors for a certain sequence X.

t:	123456789		
t mod 3:	120120120		
X:	*ATGATGACG*		$m_1\ m_2\ m_3$
A appears at:	O--O--O--	$\rightarrow$	$M_A(X) = \{3,\ 0,\ 0\}$
T appears at:	-O--O----	$\rightarrow$	$M_T(X) = \{0,\ 2,\ 0\}$
G appears at:	--O--O--O	$\rightarrow$	$M_G(X) = \{0,\ 0,\ 3\}$
C appears at:	-------O-	$\rightarrow$	$M_C(X) = \{0,\ 1,\ 0\}$

Figure 3.1 Calculation of TP vectors

Two shift operations, i.e., Left Cyclic Shift (LCS) and Right Cyclic Shift (RCS), can be applied to a TP vector $M = \{m_1, m_2, m_3\}$. The shift operations are defined to shift the values among m_1, m_2, and m_3:

$$\{m_1, m_2, m_3\} \xrightarrow{\text{LCS}} \{m_2, m_3, m_1\}$$
$$\{m_1, m_2, m_3\} \xrightarrow{\text{RCS}} \{m_3, m_1, m_2\}$$

Here the shift operations are denoted by two symbols "<<" and ">>":

$$r \text{ times LCS on } M: \quad M << r$$
$$r \text{ times RCS on } M: \quad M >> r$$

Besides, addition, subtraction, and multiplication on a TP vector M are naturally defined as:

$$\{m_1, m_2, m_3\} \pm \{m_1', m_2', m_3'\} = \{m_1 \pm m_1', m_2 \pm m_2', m_3 \pm m_3'\}$$

$$c \cdot \{m_1, m_2, m_3\} = \{cm_1, cm_2, cm_3\}$$

Here $\{m_1, m_2, m_3\}$ and $\{m'_1, m'_2, m'_3\}$ are two TP vectors and c is a real constant.

According to Equation (3-1) and Equation (3-2), a TP vector $M = \{m_1, m_2, m_3\}$ presents a Fourier spectrum by the mapping from TP vectors to complex numbers:

$$\begin{aligned} M = \{m_1, m_2, m_3\} \rightarrow U &= m_3\alpha_0 + m_1\alpha_1 + m_2\alpha_2 \\ \alpha_0 &= 1 = \cos 0 + i \sin 0 \\ \alpha_1 &= e^{-i\frac{2\pi}{3}} = \cos\frac{-2\pi}{3} + i\sin\frac{-2\pi}{3} \\ \alpha_2 &= e^{-i\frac{4\pi}{3}} = \cos\frac{-4\pi}{3} + i\sin\frac{-4\pi}{3} \end{aligned} \tag{3-3}$$

Therefore, the complex number U can be considered as a weighted summation of three components (α_0, α_1, and α_2) between each two of which there is a phase angle difference $2\pi/3$ (see Figure 3.2).

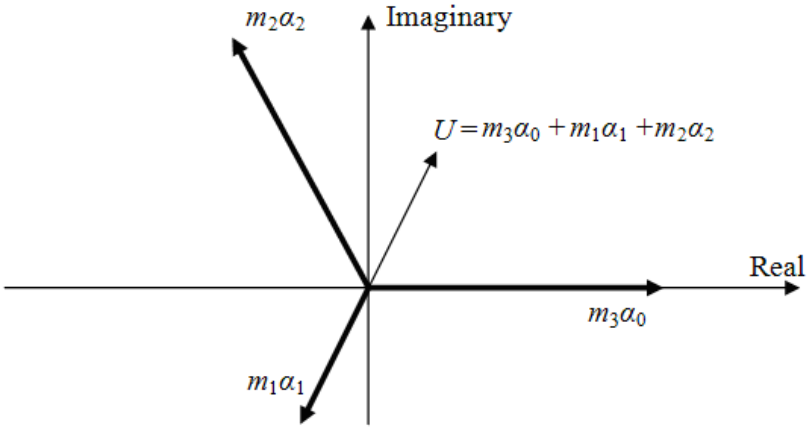


Figure 3.2 A TP vector in the complex plane

According to the definition of the operations on TP vectors, it is easy to find

that a $2\pi r/3$ counterclockwise rotation on a "TP vector mapped" complex number (a Fourier spectrum) can be easily implemented by r times LCS on its corresponding TP vector M, i.e., $M << r$; and $M >> r$ is equivalent to a $2\pi r/3$ clockwise rotation on the complex number. It is also noticed that, addition, subtraction, and multiplication on a TP vector M are equivalent to the same operations on its corresponding complex number U:

$$M_1 \pm M_2 \Leftrightarrow U_1 \pm U_2$$

$$cM \Leftrightarrow cU$$

Besides, the length of a TP vector $L(M)$ is also defined as the norm of its corresponding complex number, i.e., $L(M) = |U|$.

3.1.2 Transforming a DNA sequence into a TP sequence

A TP sequence is a new representation transformed from an original DNA sequence. It contains the information of the triplet periodicity, by calculating the TP vectors of the posterior subsequence at each position.

Firstly, the posterior subsequence of the sequence $X = \{x_t \mid t = 1, 2, \ldots, N\}$ at position t_0 is:

$$P_X(t_0) = \{x_t \mid t = t_0 + 1, t_0 + 2, ..., N\} \tag{3-4}$$

Then the TP sequence of the original DNA sequence X is defined as a sequence of TP vectors: $S(X) = \{s_t \mid t = 1, 2, \ldots, N\}$, where s_t is calculated as:

$$s_t = M_{x_t}(P_X(t)) \tag{3-5}$$

That is: for each position t, a TP vector is chosen to be s_t, from the four vectors (M_A, M_T, M_G, and M_C) of its posterior subsequence $P_X(t)$, based on the nucleotide base appearing at t, i.e., x_t. Here, an example is also given in Figure 3.3 to show how a DNA sequence is transformed into a TP sequence.

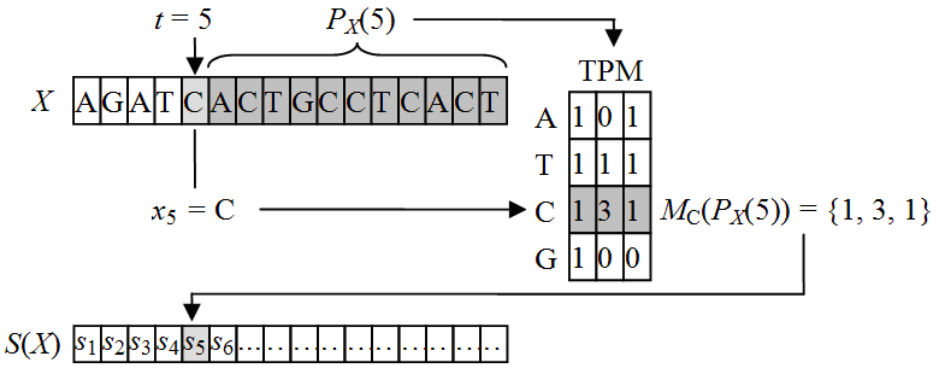


Figure 3.3 Transformation of a DNA sequence into a TP sequence

In the concept of TP sequence described above, for each position t, the posterior subsequence is considered, four TP vectors are calculated, and s_t is selected. It reveals that, with a computational complexity of $O(N^2)$, generating a TP sequence from a DNA sequence is time consuming. In order to reduce the computational complexity, a recursive scheme is developed.

It is noticed that $P_X(t+1)$ is a posterior subsequence of $P_X(t)$. Then the algorithm recursively calculates $M_\Lambda(P_X(t))$ from $M_\Lambda(P_X(t+1))$, with the following recurrence formula:

$$M_\Lambda(P_X(t)) = \begin{cases} M_\Lambda(P_X(t+1)) >> 1 & x_{t+1} \neq \Lambda \\ M_\Lambda(P_X(t+1)) >> 1 + \{1, 0, 0\} & x_{t+1} = \Lambda \end{cases} \quad (3\text{-}6)$$

The recursive process is with an initial value of $M_\Lambda(P_X(N)) = \{0, 0, 0\}$. Hence,

$M_\Lambda(P_X(t))$ can be calculated recursively from the 3' end to the 5' end. In other words, TPMs of the posterior subsequences are maintained from position N to 1. Consequently, the TP sequence can be generated by choosing s_t, at each position t, from the four vectors, i.e., $M_A(P_X(t))$, $M_T(P_X(t))$, $M_C(P_X(t))$, and $M_G(P_X(t))$. The algorithm is described in pseudocode as follows.

TP sequence generation algorithm
Input: DNA sequence $x[1\ldots N]$
Output: TP sequence $s[1\ldots N]$

```
1  For each Λ do M[Λ] = {0, 0, 0};
2  t = N;
3  While(t>0) do{
4      s[t] = M[x[t]];
5      For each Λ do{
6          M[Λ] = M[Λ] >> 1;
7          If(x[t]==Λ) M[Λ] = M[Λ] + {1, 0, 0};
8      }
9      t--;
10 }
```

An example is given in Figure 3.4. Obviously, the computational complexity is reduced to $O(N)$ by using this algorithm. This will be verified in Section 4.1.

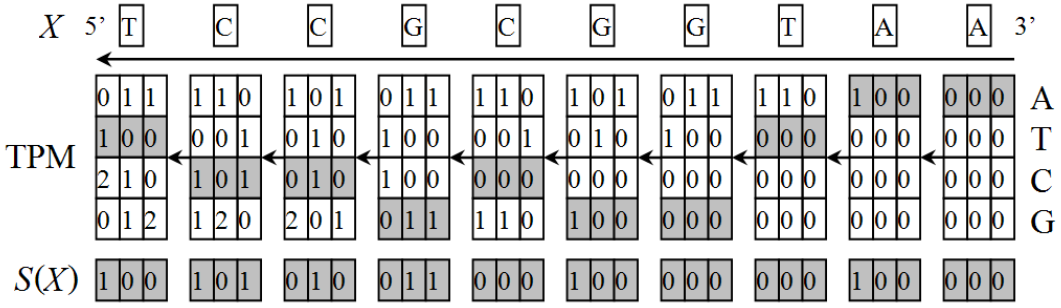


Figure 3.4 A sketch of the algorithm to generate a TP sequence

3.1.3 Generating a TP walk

A TP walk is a random walk in the TP vector's space. It starts from {0, 0, 0} and generates a moving trace according to the TP sequence $S = \{s_t \mid t = 1, 2, \ldots, N\}$. The trace can be considered as a sequence $W = \{w_t \mid t = 1, 2, \ldots, N\}$ with an initial value of $w_0 = \{0, 0, 0\}$, and for each step $t > 0$:

$$w_{t+1} = \begin{cases} w_t + \dfrac{s_t}{L(s_t)} & L(s_t) \neq 0 \\ w_t & L(s_t) = 0 \end{cases} \tag{3-7}$$

As mentioned before, any TP vector can be mapped into a complex number. So a TP sequence $\{s_t \mid t = 1, 2, \ldots, N\}$ is equivalent to a sequence of complex numbers $\{z_t \mid t = 1, 2, \ldots, N\}$, where z_t is the corresponding complex number of s_t. Therefore, Equation (3-7) reveals a corresponding definition of the TP walk in the complex plane. It is easy to find that, in the complex plane, the walk starts from the zero point 0, and for each step $t > 0$:

$$w_{t+1} = \begin{cases} w_t + \dfrac{z_t}{|z_t|} & z_t \neq 0 \\ w_t & z_t = 0 \end{cases} \tag{3-8}$$

It means to move a unit length toward the direction of z_t in the complex plane for each step. So, a DNA walk in the complex plane can be generated from any given nucleotide base sequence. This process is named Self-Adaptive Spectral Rotation (SASR) in this study. The SASR approach takes only a nucleotide base sequence as its input and provides a graphic output, i.e., the TP walk. The following section demonstrates how the SASR approach reveals coding regions and frame shifts

through its graphic output.

3.2 Behaviors of TP walks

In this section, the behavior of the SASR's output, i.e., the TP walk, is discussed under different situations, in order to demonstrate the principle by which TP walks visualize coding-related information hidden in nucleotide base sequences.

3.2.1 TP walks of simple coding and non-coding regions

Consider a simple sequence, ATGATGACG, with a high intensity of the TP property. It is easy to find that the TP sequence of this DNA sequence is S = {{0, 0, 2}, {0, 0, 1}, {0, 0, 2}, {0, 0, 1}, {0, 0, 0}, {0, 0, 1}, {0, 0, 0}, {0, 0, 0}, {0, 0, 0}}, and from Equation (3-3), its complex form should be {2, 1, 2, 1, 0, 1, 0, 0, 0}. According to the definition of TP walk, the walk trace is {1, 2, 3, 4, 4, 5, 5, 5, 5} in the complex plane, or {{0, 0, 1}, {0, 0, 2}, {0, 0, 3}, {0, 0, 4}, {0, 0, 4}, {0, 0, 5}, {0, 0, 5}, {0, 0, 5}, {0, 0, 5}} in the TP vector's space. The plot of this trace in the complex plane is shown in Figure 3.5a. It shows that the walk of this sequence has an obvious trend in the direction of the real axis.

On the other hand, when the SASR is conducted on a sequence without any TP property, the walk's behavior becomes quite different. TGTTCGACA is a randomly generated sequence with also a length of 9 and Figure 3.5b shows its TP walk trace.

No obvious trend is observed from this walk and it just randomly walks around the zero point.

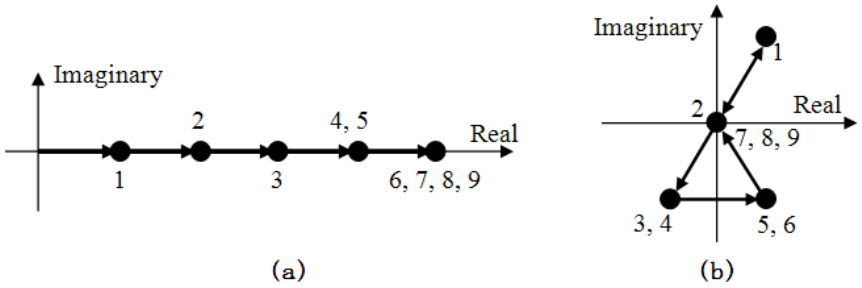


Figure 3.5 TP walks (a) Sequence ATGATGACG (b) Sequence TGTTCGACA

So a phenomenon is observed, that is: the TP walk of a sequence with a high intensity of the TP property (without any frame shift) shows an obvious trend in the direction of the real axis, while the walk of a sequence without any TP property has no such trend. This principle is reasonable and is universally satisfied for simple coding (TP-intensive) and non-coding (non-TP) sequences, because it can be proved theoretically as follows.

Consider a DNA sequence X with a TPM = $\{M_A(X), M_T(X), M_C(X), M_G(X)\}^T$. For each step t, the increment in Equation (3-7) is:

$$\frac{s_t}{L(s_t)} = \frac{M_{x_t}(P_X(t))}{L(M_{x_t}(P_X(t)))} \approx \frac{M_{x_t}(X) << t}{L(M_{x_t}(X))} \tag{3-9}$$

Here, the "approximation ≈" is employed because of the persistent distributions of nucleotide bases in triplets of coding regions and the randomness in non-coding regions (Kotlar and Lavner, 2003): In a simple coding or non-coding DNA sequence, most of the posterior subsequences share the similar entries' proportions of the TP

vectors in the TPM only with a shift caused by *t*. Meanwhile, according to Frenkel and Korotkov (2008), a certain base Λ appears at position *j* in the 3bp period with a probability:

$$\Pr\{x_t = \Lambda \text{ and } t\%3 = j\} = \frac{m_{\Lambda j}}{N} \tag{3-10}$$

Here, "%" stands for an alternative "mod" operation between two integers *a* and *b* represented as follows:

$$a\%b = \begin{cases} a \bmod b & a \bmod b \neq 0 \\ b & a \bmod b = 0 \end{cases}$$

Hence, for each step *t*, the increment of the TP walk is expected to be:

$$\begin{aligned} E\left(\frac{s_t}{L(s_t)}\right) &\approx E\left(\frac{M_{x_t}(X) << t}{L(M_{x_t}(X))}\right) = \sum_{\Lambda=A,T,C,G}\sum_{j=1}^{3}\frac{m_{\Lambda j}}{N}\cdot\frac{M_{\Lambda}(X) << j}{L(M_{\Lambda}(X))} \\ &= \sum_{\Lambda=A,T,C,G}\frac{m_{\Lambda 1}\cdot\{m_{\Lambda 2},m_{\Lambda 3},m_{\Lambda 1}\}+m_{\Lambda 2}\cdot\{m_{\Lambda 3},m_{\Lambda 1},m_{\Lambda 2}\}+m_{\Lambda 3}\cdot\{m_{\Lambda 1},m_{\Lambda 2},m_{\Lambda 3}\}}{N\cdot L(M_{\Lambda}(X))} \\ &= \sum_{\Lambda=A,T,C,G}\frac{\{m_{\Lambda 1}m_{\Lambda 2}+m_{\Lambda 2}m_{\Lambda 3}+m_{\Lambda 3}m_{\Lambda 1}, m_{\Lambda 1}m_{\Lambda 2}+m_{\Lambda 2}m_{\Lambda 3}+m_{\Lambda 3}m_{\Lambda 1}, {m_{\Lambda 1}}^2+{m_{\Lambda 2}}^2+{m_{\Lambda 3}}^2\}}{N\cdot L(M_{\Lambda}(X))} \end{aligned} \tag{3-11}$$

Kotlar and Lavner's work (2003) shows that the entries in the TP vectors are biased in a simple coding sequence and uniformly random in a non-coding sequence. That is:

Simple coding sequence: $m_{\Lambda 1} \neq m_{\Lambda 2} \neq m_{\Lambda 3}$

Non-coding sequence: $m_{\Lambda 1} \approx m_{\Lambda 2} \approx m_{\Lambda 3}$

Then we have:

Simple coding sequence: ${m_{\Lambda 1}}^2 + {m_{\Lambda 2}}^2 + {m_{\Lambda 3}}^2 > m_{\Lambda 1}m_{\Lambda 2} + m_{\Lambda 2}m_{\Lambda 3} + m_{\Lambda 3}m_{\Lambda 1}$

Non-coding sequence: ${m_{\Lambda 1}}^2 + {m_{\Lambda 2}}^2 + {m_{\Lambda 3}}^2 \approx m_{\Lambda 1}m_{\Lambda 2} + m_{\Lambda 2}m_{\Lambda 3} + m_{\Lambda 3}m_{\Lambda 1}$

It shows that, in Equation (3-11), for a simple coding sequence, the third element of the expected increment dominates the other two. According to Equation (3-3), the TP walk of this coding sequence should move rightward in the complex plane. On the other hand, since the three elements of the vector are balanced for a non-coding sequence, the walk of this non-coding sequence should appear to be random around the zero point. This principle will be verified with real DNA data in Section 4.2.1.

3.2.2 TP walks of C_0-I-C_1 chains

According to Section 3.2.1, the TP walk of the sequence from a single coding region shows an obvious trend in the direction of the real axis. However, the TP walk of a long DNA sequence, which consists of a number of coding and non-coding regions, behaves more complicatedly. Consider a short chain which consists of a coding region, a non-coding region, and a coding region sequentially from the 5' end to the 3' end. It is denoted as C_0-I-C_1 as shown in Figure 3.6. The two coding regions C_0 and C_1 are from a same organism, therefore share a same TP profile. Because the visualization focuses on the coding regions' general locations and the frame shifts, rather than exact boundaries, it is safe to assume that the lengths of C_0 and C_1 are multiples of 3, excluding the incomplete codons. Therefore, the non-coding region I between them indicates a frame shift caused by insertions or deletions. The

difference between the coding regions' reading directions indicates a frame shift caused by an inversion.

According to the definition of TP sequence, which only takes the posterior subsequence into consideration, the walk in C_1 is not influenced by I or C_0, and it will be in the positive real direction as usual. However, the walks in C_0 and I are influenced by their posterior parts, which are I-C_1 and C_1 respectively. Firstly, the walk in I, which is influenced by the posterior part C_1, is taken into consideration. Suppose that a nucleotide base Λ appears at position t in the I part (Figure 3.6), thus we have:

$$s_t = M_\Lambda(P_{I-C_1}(t)) = M_\Lambda(P_I(t)) + M_\Lambda(C_1) >> (N_I - t) \tag{3-12}$$

Here N_I means the length of the non-coding region I. According to the discussion in Section 3.2.1, the first term in the right hand side $M_\Lambda(P_I(t))$ is a complex random variable with an expected value of 0, since I has no TP property and the three elements in this vector are balanced. Meanwhile $M_\Lambda(C_1)$ is a non-zero constant vector, since C_1 is a coding region with the TP property. However, because I is non-coding, the position where base Λ appears is uniformly random (Kotlar and Lavner, 2003). It means that t is uniformly random and so is $N_I - t$. Therefore, the second term $M_\Lambda(C_1) >> (N_I - t)$ should be also random with an expected value of 0. So, in view of the total effect, the walk in this part should be random around a relatively stable point.

For the TP walk in C_0 part, it is also supposed that a base Λ appears at position t in C_0 part (Figure 3.6). Thus we have:

$$\begin{aligned} s_t &= M_\Lambda(P_{C_0\text{-}I\text{-}C_1}(t)) \\ &= M_\Lambda(P_{C_0}(t)) + M_\Lambda(I) >> (N_{C_0} - t) + M_\Lambda(C_1) >> (N_I + N_{C_0} - t) \\ &= M_\Lambda(P_{C_0}(t)) + M_\Lambda(I) << t + (M_\Lambda(C_1) << t) >> N_I \qquad (N_{C_0} \bmod 3 = 0) \end{aligned} \tag{3-13}$$

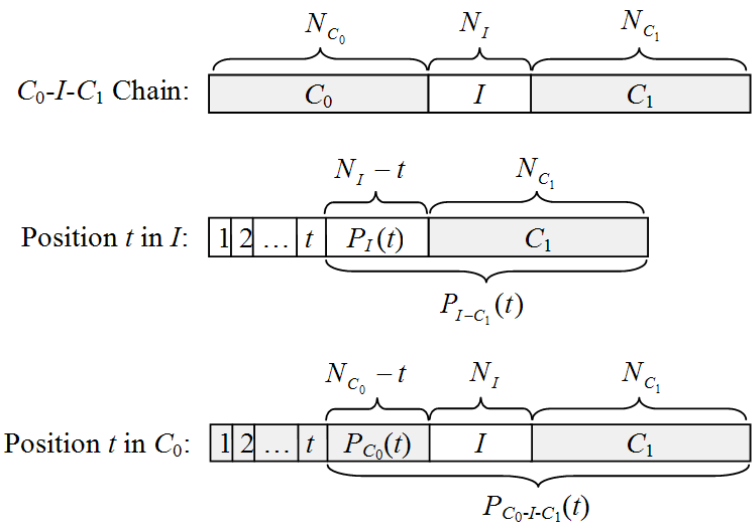


Figure 3.6 A C_0-I-C_1 chain

Obviously, the first term in Equation (3-13) just indicates the original behavior of the TP walk in C_0 without the influence from I and C_1, and it is expected to be in the positive real direction as mentioned previously. The second term is nearly 0, since there is no dominant element in $M_\Lambda(I)$.

Now focus is given to the third term in Equation (3-13). The difference between the coding regions' reading directions reveals a frame shift caused by an inversion. There are two reading directions, i.e., the forward and reverse directions (Zhang and Zhang, 1994) for each region. Hence, the chain can be coherent (C_0 and

C_1 are in a same reading order) or incoherent (C_0 and C_1 are in different reading orders). There is no harm to assume that C_1 is in the forward direction, because if not, the following derivation is similar. It is noticed that, according to Kotlar and Lavner's work (2003), for a certain organism, nucleotide base Λ has its preferred triplet position r_Λ (a real number in (0, 3] as an expected value) in the 3bp period. It causes $M_\Lambda(C_1)$ to be with an expected phase angle of $-2\pi r_\Lambda/3$ in the complex plane. Then the behavior of the TP walk in C_0 is discussed in the following two cases.

If the chain is coherent (C_0 is also in the forward direction), the preferred position of Λ in C_0 is r_Λ. Since t is just a position where Λ appears, in view of the total effect, $M_\Lambda(C_1) << t$ is likely to cause the same effect as what $M_\Lambda(C_1) << r_\Lambda$ does. It means a $2\pi r_\Lambda/3$ counterclockwise rotation on $M_\Lambda(C_1)$, which is with an expected phase angle of $-2\pi r_\Lambda/3$, and the production is with an expected phase angle of 0. In other words, $M_\Lambda(C_1) << t$ is expected to be a positive real number. Then the direction of the third term only depends on the length of I, i.e., N_I. The frame shift between the two coding regions (without inversion) is $\Delta = N_I \bmod 3$. It is easy to find that: if $\Delta = 0$, the walk in C_0 will still be in the positive real direction, which is the same direction as in C_1. Otherwise, there will be a corner between these two coding regions, and the walk trace in C_0 will be an arc since the first term in Equation (3-13), $M_\Lambda(P_{C_0}(t))$, becomes weaker and weaker with the growth of t, until the third term totally dominates in s_t at the end of C_0, where the walk direction should only depend

on the value of Δ. Accordingly, there is a strong relationship between Δ and the corner's shape, which is called here the "corner rule" (Figure 3.7). When the chain is coherent (C_0 and C_1 are in a same reading direction), the change of the walk direction on the corner depends on Δ:

Δ = 0 : The direction remains the same (goes straight)

Δ = 1 : The direction rotates $2\pi/3$ counterclockwise (turns left)

Δ = 2 : The direction rotates $2\pi/3$ clockwise (turns right)

$$s_t|_{C_0} = M_\Lambda(P_{C_0}(t)) + (M_\Lambda(C_1) << t) >> \Delta$$
$$s_t|_{\text{end of } C_0} = (M_\Lambda(C_1) << t) >> \Delta$$

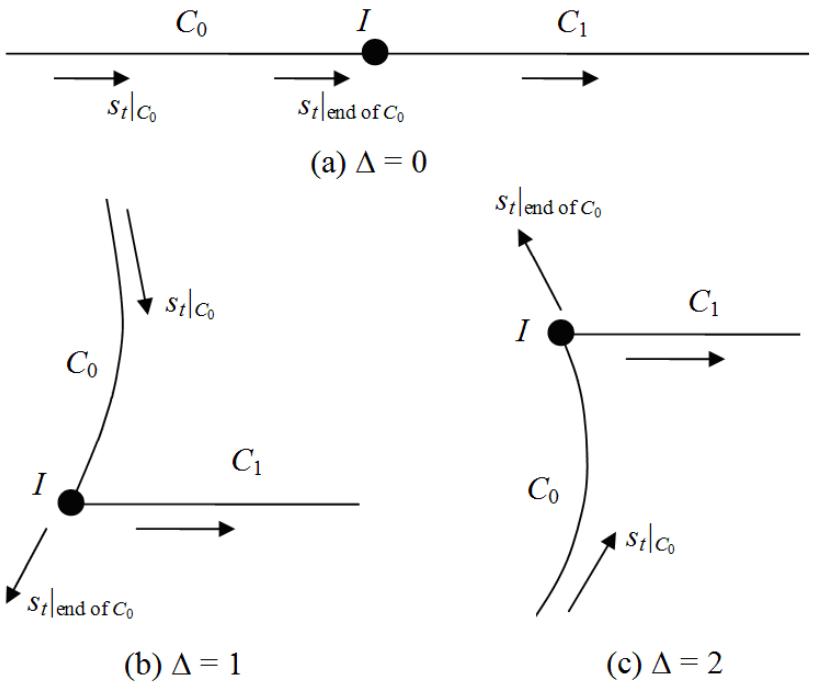


Figure 3.7 A sketch of the TP walk trace of a coherent C_0-I-C_1 chain

This corner rule is satisfied only if the chain is coherent. Section 4.2.2 will

verify this rule using real DNA data. On the other hand, if the chain is incoherent (C_0 is in the reverse reading direction, different from that of C_1), the triplets in C_0 are read from the complementary strand in the reverse direction (Zhang and Zhang, 1994). So the preferred position of Λ in C_0 turns to be $4 - r_{\Lambda'}$ (as the mirror image of $r_{\Lambda'}$ with the symmetry center 2). Here Λ' denotes the complementary base of Λ, i.e., A′ = T, T′ = A, G′ = C, and C′ = G. Then in view of the total effect, $M_\Lambda(C_1) << t$ is the same as $M_\Lambda(C_1) << (4 - r_{\Lambda'}) = M_\Lambda(C_1) << (1 - r_{\Lambda'})$. It means a $2\pi(1 - r_{\Lambda'})/3$ counterclockwise rotation on $M_\Lambda(C_1)$ and the production is with an expected phase angle of $2\pi(1 - r_\Lambda - r_{\Lambda'})/3$. In view of the total effect, the third term $(M_\Lambda(C_1) << t) >> N_I$ has an expected direction. It reveals that the walk in C_0 part has its trend and the walk trace also follows an arc. But the expected direction depends on some statistics of the organism besides a simple Δ value, including the proportions and the preferred triplet positions of the nucleotide bases.

3.2.3 TP walk of a complete DNA sequence

A complete DNA sequence is a long chain, which is *I*-*C*-*I*-*C*-…-*C*-*I*. It is easy to find that the behavior of its TP walk shows accumulative effect of its short *C*-*I*-*C* sub-chains from the 3’ end to the 5’ end. Therefore, the TP walk of a complete sequence should follow the rules:

(1) The walk traces of the coding regions are arcs and the walk in the last

coding region is in the direction of the real axis.

(2) The walk in the non-coding regions is always random and moves around stable points.

(3) If two neighboring coding regions are in a same reading order (the local C_0-I-C_1 chain is coherent), the shape of the corner between them follows the corner rule described in Section 3.2.2.

The rule (3) is proved as follows.

The coding regions from the 5' end to the 3' end are numbered as C_0, C_1, …, C_{K-1}, C_K. Consider the corner between two neighboring coding regions C_{k-1} and C_k, which are both in the forward direction (the situation is similar if they are both in the reverse direction). Suppose that a base Λ appears at position t^- in C_{k-1} and t^+ in C_k, and these two positions are close to the corner (t^- and t^+ are local index numbers for C_{k-1} and C_k, namely, t^- is nearly the length of C_{k-1} and t^+ is close to 0). Then calculations of the expected walk directions are made at these two positions.

Ignore the influence from the inner I parts since it is nearly 0 as discussed previously, and also ignore the very short posterior subsequence at t^- in C_{k-1} since position t^- is close to the corner (the end of C_{k-1}). Then we have:

$$
\begin{aligned}
s_{t^-} &\approx \sum_{i=k}^{K} (M_\Lambda(C_i) << t^-) >> \Delta(C_{k-1}, C_i) \\
&\approx \sum_{i=k}^{K} (M_\Lambda(C_i) << r_\Lambda) >> \Delta(C_{k-1}, C_i) \qquad (r_\Lambda \text{ is the preferred triplet position of } \Lambda) \\
&= \left[\sum_{i=k}^{K} (M_\Lambda(C_i) << r_\Lambda) >> \Delta(C_k, C_i) \right] >> \Delta(C_{k-1}, C_k)
\end{aligned}
$$

$$s_{t^+} \approx M_\Lambda(P_{C_k}(t^+)) + \sum_{i=k+1}^{K}(M_\Lambda(C_i) << t^+) >> \Delta(C_k, C_i)$$
$$\approx M_\Lambda(P_{C_k}(t^+)) + \sum_{i=k+1}^{K}(M_\Lambda(C_i) << r_\Lambda) >> \Delta(C_k, C_i)$$

Since t^+ is close to the corner (the start of C_k), the posterior subsequence at t^+ in C_k is nearly the entire C_k with a shift, that is:

$$M_\Lambda(P_{C_k}(t^+)) \approx M_\Lambda(C_k) << t^+ \approx M_\Lambda(C_k) << r_\Lambda$$

Accordingly,

$$s_{t^+} \approx M_\Lambda(C_k) << r_\Lambda + \sum_{i=k+1}^{K}(M_\Lambda(C_i) << r_\Lambda) >> \Delta(C_k, C_i)$$
$$= \sum_{i=k}^{K}(M_\Lambda(C_i) << r_\Lambda) >> \Delta(C_k, C_i)$$

$$\therefore s_{t^-} = s_{t^+} >> \Delta(C_{k-1}, C_k)$$

It reveals that the walk direction rotates on the corner depending on $\Delta(C_{k\text{-}1}, C_k)$.

It is also noticed that, in such a complete DNA sequence, since a single coding region is much shorter than the summation of its posterior coding regions, $M_\Lambda(P_{C_0}(t))$ becomes very weak compared with the influence from the posterior coding regions. Therefore, in most cases, the arcs are with weak bends, so the TP walk trace of the complete DNA sequence can be considered as a polyline.

Figure 3.8 shows the TP walk trace of the complete *Homo sapiens* (Human) mitochondrial DNA sequence (GenBank no. NC_001807). The total 13 coding regions are marked in different colors, and the Δ value between each two of them is shown as well. It is clear in the figure that the coding regions stay on the arcs while

the non-coding regions stay on the corners or around relatively stable points. The corner between each adjacent two of the first 11 coding regions follows the corner rule. The corner rule is not applicable to the two corners among the 11th, the 12th, and the 13th coding regions, because the 12th coding region is in the reverse reading direction. However, the 12th coding region also stays on an arc. Meanwhile, the top-right of Figure 3.8 shows that the curves of the real part and the imaginary part fluctuate with the alternation of the coding regions and stay relatively constant in the non-coding regions. An experiment in Section 4.2.3 will show the walks of more real DNA sequences and numerically verify the corner rule in complete sequences.

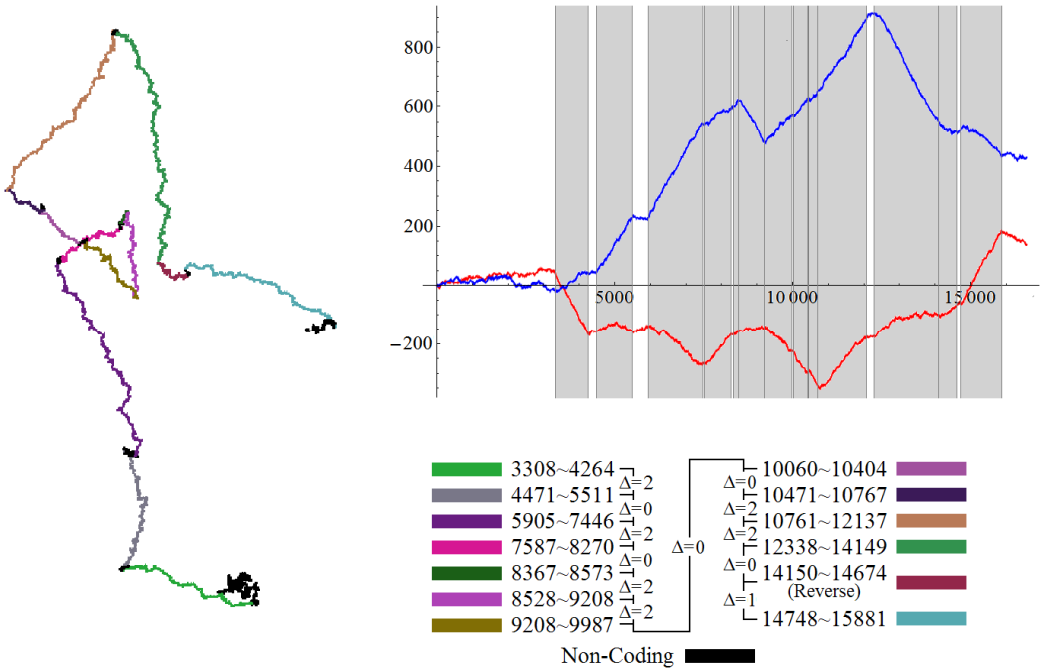


Figure 3.8 The TP walk trace of the complete *Homo sapiens* (Human) mitochondrial DNA sequence in the complex plane with the coding regions marked in different colors. The top-right is the plot of the real part (red) and the imaginary part (blue) against the position value *t*, and the dark areas stand for the coding regions.

3.3 Computational analyses of TP walks

After the main process of the SASR, a TP walk visually discriminates coding regions from non-coding regions and also reveals frame shifts. However, as a visualization method, the SASR approach does not provide exact numerical results for the prediction. So, some computational analyses of TP walks are proposed in this section for various purposes of gene analysis.

3.3.1 Rightward Rate (RR) measure

As mentioned in Section 3.2.1, the TP walk of a simple coding sequence has an obvious trend to move rightward and the TP walk of a non-coding one moves randomly around the zero point. To quantitatively verify this principle in practice, it is significant to quantitatively investigate to what extent a TP walk moving rightward. For this purpose, a Rightward Rate (RR) measure is presented here. For a given DNA sequence, an RR measure is calculated from its TP walk $W = \{w_t \mid t = 0, 1, 2, \ldots, N\}$:

$$RR = \frac{1}{N} \max\{\mathrm{Re}(w_t) \mid t = 1, 2, ..., N\} \tag{3-14}$$

Here, Re(w) stands for the real part of the complex number w. This measure reveals the average speed at which the walk moves rightward (in the positive real direction) in the complex plane. In Section 4.2.1, the RR measure is used to quantitatively investigate, to what extent, the TP walks are different for simple coding sequences and non-coding sequences.

According to the above definition, an RR measure should be not less than 0 and does not allow revealing the walk trend that to move leftward. However, in some cases, a leftward trend should also be considered (details in Section 5.2). So a Symmetrical Rightward Rate (SRR) is presented as:

$$SRR = \frac{1}{N}\left[\max\{\mathrm{Re}(w_t) \mid t = 1, 2, ..., N\} + \min\{\mathrm{Re}(w_t) \mid t = 1, 2, ..., N\}\right] \qquad (3\text{-}15)$$

If a walk has an obvious trend to move rightward, its SRR measure tends to be positive, while a walk to move leftward provides a negative SRR measure. And a walk to move randomly around the zero point has an SRR measure close to 0.

3.3.2 Fixed Scale Numerical Differentiation (FSND) analysis

The RR measure is developed only to reveal the average speed at which a walk moves rightward. It can be further used in the classification of simple coding and non-coding sequences, but could hardly be an ideal numerical solution to coding region prediction, since it ignores local patterns in more complicated walks. However, for complete DNA sequences, a straightforward method named Fixed Scale Numerical Differentiation (FSND) can be developed to investigate the local moving speed of the walk so that the locations of coding regions can be revealed.

Consider a DNA sequence from chromosome 2 of *S. pombe* (GenBank no. NC_003423), ranging from position 432,001 to 434,800. This sequence includes the gene *SPBC*582.08 containing 3 exons (see Table 3.1), which was used as a typical

gene sequence in Kotlar and Lavner's work (2003). After applying the SASR to this sequence, the output TP walk $W = \{w_t \mid t = 0, 1, 2, \ldots, N\}$ is shown in Figure 3.9a. Then a Fixed Scale Numerical Differentiation (FSND) is conducted on this output in order to investigate the local moving speed v_t at each position t in the walk. With a fixed analysis scale Δt (as an example, $\Delta t = 127$ is used here), the indicator v_t is calculated following:

$$\Delta x = \mathrm{Re}(w_{t+\Delta t}) - \mathrm{Re}(w_{t-\Delta t}) \qquad \Delta y = \mathrm{Im}(w_{t+\Delta t}) - \mathrm{Im}(w_{t-\Delta t})$$
$$v_t = \sqrt{\Delta x^2 + \Delta y^2} \tag{3-16}$$

Here functions $\mathrm{Re}(w)$ and $\mathrm{Im}(w)$ are the real part and the imaginary part of the complex number w respectively. The local moving speed v_t is plotted in Figure 3.9b. As expected, Figure 3.9b shows that the three exons, which are indicted by the three thick bars, raise three peaks of the local moving speed v_t. After selecting a suitable threshold v_0, each position t can be assigned as "coding" or "non-coding" by checking whether $v_t \geq v_0$. For this sample case, after trying all possible values of v_0, $v_0 = 0.069$ is found to make the prediction achieve optimal performance, with an accuracy of 92%. In this context, accuracy (Ac), sensitivity (Sn), and specificity (Sp) are calculated as follows:

$$\mathrm{Ac} = \frac{\text{number of correctly assigned sites}}{\text{length of the sequence}}$$
$$\mathrm{Sn} = \frac{\text{number of correctly assigned coding sites}}{\text{number of coding sites}} \tag{3-17}$$
$$\mathrm{Sp} = \frac{\text{number of correctly assigned non-coding sites}}{\text{number of non-coding sites}}$$

Table 3.1 The exons in the gene *SPBC*582.08 from chromosome 2 of *S. pombe*

Exon	Start base	End base	Length	Reading frame
1	432,550	433,012	463	1
2	433,204	434,003	800	3
3	434,057	434,311	255	2

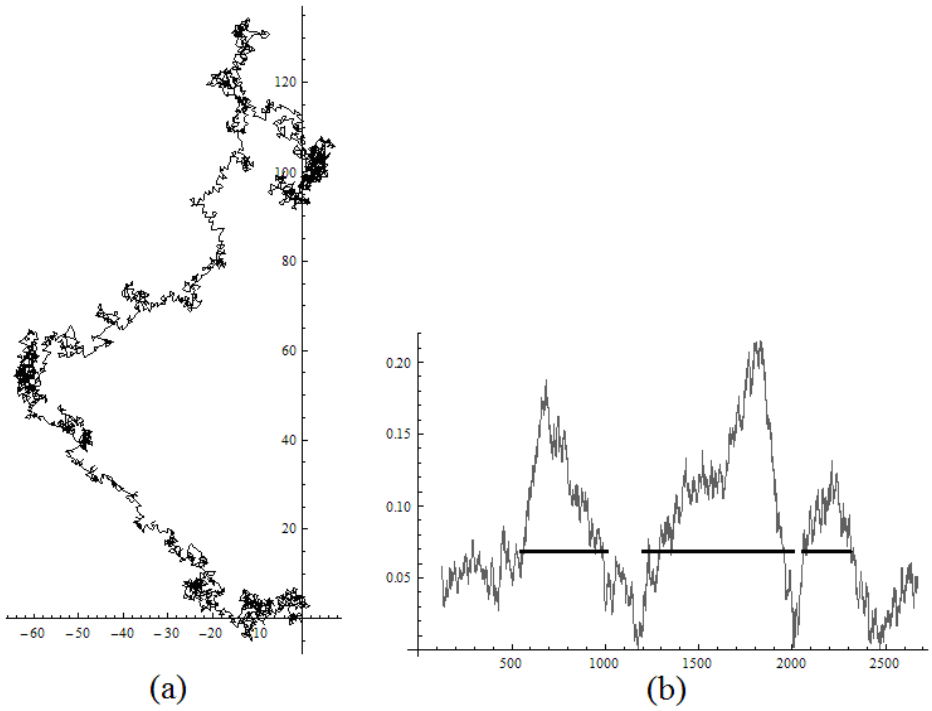


Figure 3.9 The numerical differentiation on the TP walk of the gene *SPBC*582.08 in chromosome 2 of *S. pombe* (a) The TP walk (b) Plot of the local moving speed v_t

For an unannotated sequence, in order to select the best moving speed threshold v_0, which optimizes the accuracy of the prediction, a simple training is necessary. A process series, consisting of an SASR step followed by an FSND step (denoted as "SASR-FSND"), is firstly applied to some annotated sequences (training set) from the relatives of the query sequence. For each sequence in the training set, the best

threshold v_{opt} is found to optimize the accuracy of the prediction, and meanwhile the average value of v_t, i.e., v_{avg}, can be calculated over all positions t. A regression model can be built to describe the relationship between v_{avg} and v_{opt}. Therefore, the best moving speed threshold for the query sequence can be estimated by using this regression model. The experiment in Section 4.3.1 will show more details with real DNA data.

3.3.3 T-Z-T analysis

In the above section, the FSND assigns each site in a sequence as coding or non-coding, so that coding region prediction can be made. However, the sites are not organized in the form of regions, but considered individually. And the local moving speed is the only concern in the FSND. It causes that the FSND takes less advantage of the "corner rule" in a TP walk to understand the frame shifts between regions. And moreover, it is also training dependent and with a fixed analysis scale. Therefore, a more flexible algorithm should be developed instead.

The TP walk of a DNA sequence can be viewed as an accumulative local stationary process (Bernaola-Galvan et al., 2001; Hamilton, 2008), in which the complex increment $\Delta w_t = z_t / |z_t|$ has different expected values and variances in different regions. An efficient algorithm was proposed (Bernaola-Galvan et al., 2001) to deal with the segmentation of such a local stationary process, which detects the

moving trend in the walk, ignoring small local fluctuations (Vaglica et al., 2008). Here, this algorithm is adopted to find reasonable partitions of a TP walk. For each step t_0 in a TP walk, the sample means of all complex increments before and after it are calculated as follows:

$$\overline{\Delta w}_{\text{Before}}(t_0) = \frac{\sum_{t=1}^{t_0} \Delta w_t}{t_0} = \frac{w_{t_0}}{t_0} \quad \text{and} \quad \overline{\Delta w}_{\text{After}}(t_0) = \frac{\sum_{t=t_0+1}^{N} \Delta w_t}{N - t_0} = \frac{w_N - w_{t_0}}{N - t_0}$$

And the sample standard deviations:

$$S^2{}_{\text{Before}}(t_0) = \frac{\sum_{t=1}^{t_0} \left| \Delta w_t - \overline{\Delta w}_{\text{Before}}(t_0) \right|^2}{t_0 - 1} = \frac{t_0 (1 - \left| \overline{\Delta w}_{\text{Before}}(t_0) \right|^2)}{t_0 - 1}$$

$$S^2{}_{\text{After}}(t_0) = \frac{\sum_{t=t_0+1}^{N} \left| \Delta w_t - \overline{\Delta w}_{\text{After}}(t_0) \right|^2}{N - t_0 - 1} = \frac{(N - t_0)(1 - \left| \overline{\Delta w}_{\text{After}}(t_0) \right|^2)}{N - t_0 - 1}$$

Therefore, the two-sample t statistic is:

$$T(t_0) = \frac{\left| \overline{\Delta w}_{\text{Before}}(t_0) - \overline{\Delta w}_{\text{After}}(t_0) \right|}{\sqrt{\frac{(t_0 - 1) S^2{}_{\text{Before}}(t_0) + (N - t_0 - 1) S^2{}_{\text{After}}(t_0)}{N - 2} \cdot \left(\frac{1}{t_0} + \frac{1}{N - t_0} \right)}}$$

$$= \left| \overline{\Delta w}_{\text{Before}}(t_0) - \overline{\Delta w}_{\text{After}}(t_0) \right| \cdot \sqrt{\frac{t_0 (N - t_0)(N - 2)}{t_0 N (1 - \left| \overline{\Delta w}_{\text{Before}}(t_0) \right|^2) + N (N - t_0)(1 - \left| \overline{\Delta w}_{\text{After}}(t_0) \right|^2)}}$$

(3-18)

And then, the step t_0, which has the maximum t statistic $T(t_0)$, is found. Since the t statistic T follows the t-distribution with a degree of freedom N–2, the significance (or 1–p-value) of $T(t_0)$ can be calculated from the Probability Density Function (PDF) of the t-distribution (Figure 3.10a):

$$\text{(significance) } 1 - p(t_0) = \Pr\{T \le T(t_0)\} = \int_{\tau=-T(t_0)}^{T(t_0)} f_{N-2}(\tau) d\tau$$

Here, the function $f_{N-2}(\tau)$ is the PDF of the *t*-distribution with a degree of freedom N–2. The whole walk is cut into two segments if the significance is beyond a certain level P_0 (95% or 99% as commonly used in many *t*-test applications). The algorithm then continues recursively on the two segments created by the cut until no new cut can be made. It is important to note that, before a new cut is accepted, we also compute the *t* statistic between the right new segment and its right neighbor (created by a previous cut), as well as the *t* statistic between the left new segment and its left neighbor, and check whether both the statistics have significance exceeding P_0. After this recursive procedure, the whole walk is said to be "segmented at significance level P_0" (Bernaola-Galvan et al., 2001).

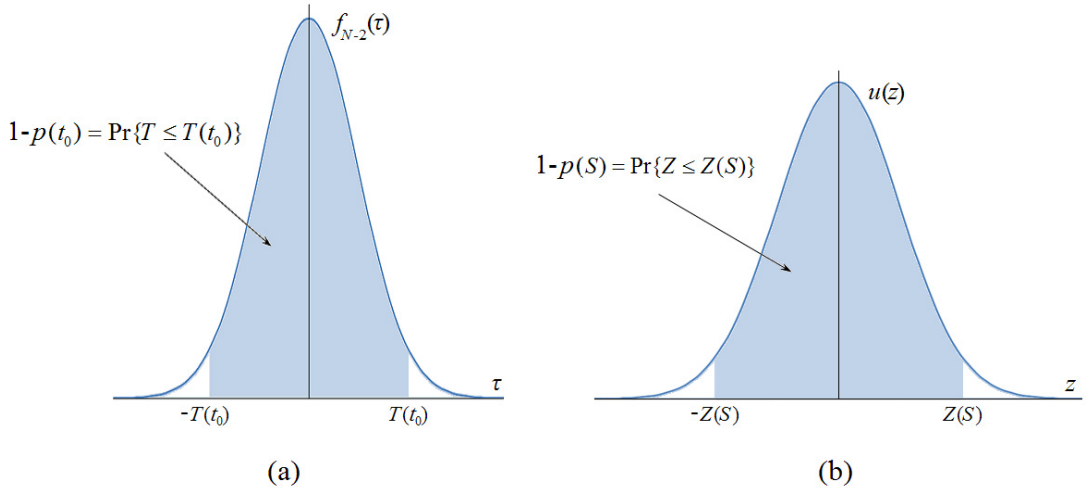


Figure 3.10 Calculation of significance (a) for the *t*-statistic (b) for the *z*-statistic

To provide coding region candidates in a long DNA sequence, after obtaining

the TP walk by the SASR, the walk is segmented at significance level $P_0 = 95\%$ (see more discussions about this significance level in the end of this section). Since coding and non-coding regions have different patterns (regime-switching) in a TP walk, it is proposed to separate the coding regions from the non-coding ones by this segmentation (Figure 3.11a). And then, the non-coding segments are filter out by conducting a *z*-test as below.

At each step t in a "hypothetical" non-coding segment S with a length of l, the increment Δw_t is a complex random variable nearly uniformly distributed on the unit circle. It has an expected value of 0 and a variance of 1. Therefore, the mean value of all l increments in this segment (i.i.d. random variables) follows the normal distribution with an expected value of 0 and a variance of $1/l$, and

$$\overline{\Delta w}\sqrt{l} \sim N(0,1)$$

Here $N(0,1)$ is the standard normal distribution in the complex plane. Therefore, a z statistic of the given segment S is set:

$$Z(S) = \left|\overline{\Delta w}\sqrt{l}\right| = \frac{\left|w_{\text{end}} - w_{\text{begin}-1}\right|}{\sqrt{l}} \tag{3-19}$$

The significance of $Z(S)$ can be calculated from the PDF of the standard normal distribution (Figure 3.10b):

$$\text{(significance)}\quad 1\text{-}p(S) = \Pr\{Z \le Z(S)\} = \int_{z=-Z(S)}^{Z(S)} u(z)dz$$

Here, function $u(z)$ is the PDF of the standard normal distribution. If the z statistic of a segment has significance below a certain level P_1, this segment is

filtered out as a non-coding one. As shown in Figure 3.11a, the significances of the z statistics of the segments appear to be "polarized": the coding segments have significance close to 1, the non-coding ones have much lower values, and the changes between them are not continuous. Therefore, for this individual case, using any significance level P_1 in a great range will get a same filtering result, and here P_1 as 85% is chosen (see more discussions about this significance level in the end of this section).

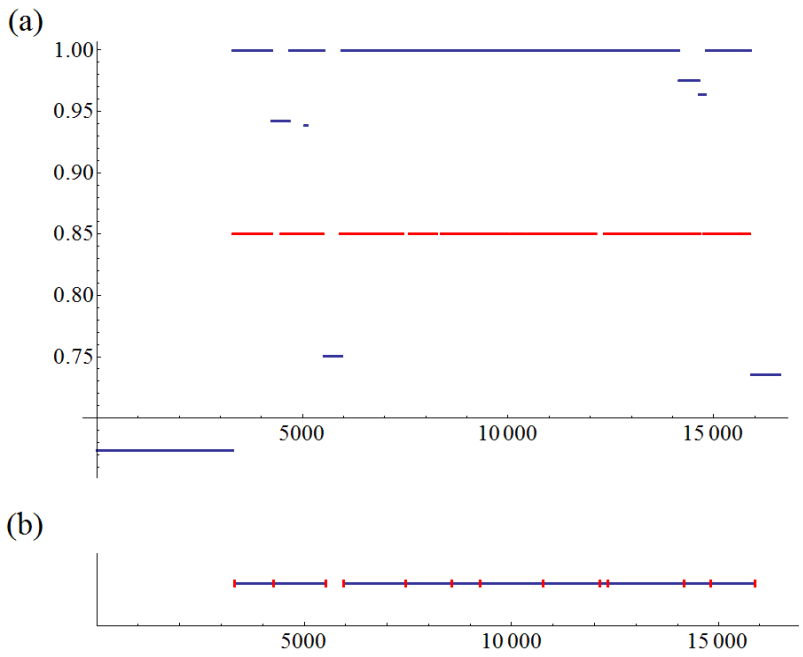


Figure 3.11 T-Z-T analysis on the TP walk of the mtDNA from *Homo sapiens* (a) The segments after the first round segmentation at significance level 95%. The horizontal axis represents the sequence position, the vertical axis indicates the significances of the z statistics of the segments, the blue bars show the segmentation result, and the red bars show the true coding regions. (b) The coding region candidates (blue bars) provided by the proposed procedure. The red vertical bars show the boundaries of the regions.

The previous segmentation aims at separating the coding regions from the non-coding ones, and it may also cut a potential coding region into pieces (Figure 3.11a). Hence, after filtering out the non-coding segments, consecutive coding segments, which are not interrupted by non-coding ones, are combined to obtain some long segments. For each long segment, segmentation is conducted at another significance level P_2, to separate the coding regions with frame shifts. The frame shifts will change the moving directions (trends) in a TP walk (Figure 3.8) and the changes are greater than those between coding (moving pattern) and non-coding (stable pattern) regions. Hence, in general, $P_2 > P_0$ is needed. Here, $P_2 = 99\%$ is chosen (see more discussions about this significance level in the end of this section). Figure 3.11b shows the segments obtained after this procedure, which consists of two "*t*-test segmentation" and a "*z*-test filter", called the T-Z-T analysis. It provides a set of coding region candidates as a prediction. Each candidate represents the rough location of a single coding region, or a local C_0-I-C_1 chain with very short non-coding region I and $\Delta = 0$ (no frame shift), which needs some other methods and extra information to further break it down. In Section 4.3.2, this prediction will be evaluated by comparisons with the true experimental result.

According to the description of the T-Z-T process, after the first *t*-test segmentation, the TP walk of a sequence is organized in partitions between every two "pattern change positions". Instead of using a fixed analysis scale, a significance

level P_0 is involved to control the analysis scale, by setting the expected sensitivity to the "pattern change" in the TP walk. A lower value of P_0 causes higher sensitivity to the pattern change, more cuts are made in the first *t*-test segmentation, and more details in the graph concerned. In general, P_0 = 95% is found to be appropriate to detect the pattern change between fast moving with a trend (coding pattern) and the uniformly random walk (non-coding pattern). So, P_0 = 95% is set to separate the coding and non-coding regions in the first *t*-test segmentation. For the *z*-test filter, it is easy to find that, the significance level P_1 actually indicates the expected specificity of the prediction, i.e., the rate of the number of correctly filtered out non-coding regions to the number of real non-coding regions. Therefore, in the application, P_1 can be directly set to the expected specificity, rather than obtained by training, compared with the threshold v_{opt} in the FSND. And moreover, the third significance level P_2 is set up to control the expected sensitivity to the "direction change" between connected coding regions in the TP walk. Similar to P_0, a lower value of P_2 makes it more sensitive to the change of the trend direction. In general, P_2 = 99% is found to be suitable to detect the $\pm 2\pi/3$ direction change caused by a frame shift while ignoring other slighter ones. Accordingly, the T-Z-T analysis uses three significance levels P_0, P_1, and P_2 to control the analysis scale, the expected specificity of the prediction, and the usage of the corner rule, respectively. The configuration is more flexible and no training process is involved.

3.4 Summary

In this chapter, a triplet vector called TP vector has been given to extract the triplet periodicity from sequences. Using the vector as a basic unit, a new representation of DNA, the TP sequence, has been proposed as well as its efficient generating algorithm with a computational complexity of $O(N)$. An accumulative effect can be observed in the TP walk, which is a random walk generated from the TP sequence. It shows that the walk has some special patterns, visually revealing the locations of the coding and non-coding regions in a DNA sequence, as well as the frame shifts. Besides such a visualization approach, namely, the SASR approach, for various analysis purposes, some operable computational methods have been developed based on the SASR's result, including the RR measure, the FSND approach, and the T-Z-T approach. Among them, the FSND approach and the T-Z-T approach can extract numerical results from a TP walk and provide computational predictions of coding regions.

CHAPTER 4

VERIFICATIONS OF THE SASR METHOD

The mathematical basis and the algorithm of the new SASR approach have been introduced in Chapter 3. This chapter will verify this new approach by applying it to real DNA data. Experiments will be conducted to test the computational time of the new approach, show its practical behaviors, and evaluate the performance of the computational coding region prediction by using this approach.

4.1 Computational time of the SASR algorithm

To test the computational time in generating a TP sequence from a given DNA sequence, a simple program is written in the C++ language and executed on a personal computer with Xeon(TM) CPU 2.8GHz and 2.0GB memory. It randomly generates 1,000 artificial DNA sequences with random lengths (ranging from 20,000bp to 20,000,000bp) and transforms them into TP sequences by using the SASR algorithm in Section 3.1.2. The computational time of the 1,000 transformations are recorded and plotted in Figure 4.1. It shows that the practical computational time rises from nearly 0ms to 4,200ms with the sequence's length N increasing from 20,000bp to 20,000,000bp linearly. It reveals that the practical computational complexity is $O(N)$.

According to Section 3.1.2, the computational complexity of the transformation from a TP sequence to a TP walk is obviously linear, so the computational complexity of the entire SASR algorithm is $O(N)$, operable in dealing with a great amount of DNA sequence data.

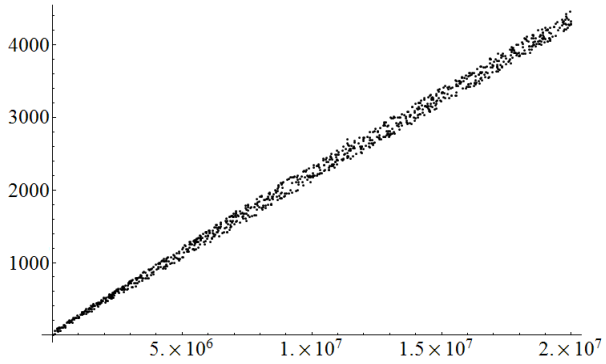


Figure 4.1 Plot of the computational time against the sequence's length N. The horizontal axis stands for the sequence's length and the vertical axis stands for the computational time in millisecond.

4.2 Visual patterns in TP walks of real DNA sequences

In this section, C++ programs are made to test practical behaviors of TP walks of real DNA sequences, under different situations, including cases for simple coding/non-coding sequences, C_0-I-C_1 chains, and complete sequences.

4.2.1 TP walks of simple coding and non-coding sequences

Section 3.2.1 shows that the TP walk of a simple coding sequence has an obvious trend in the direction of the real axis, while a non-coding sequence (with no

TP property) shows no such trend. Besides the artificial examples in Section 3.2.1, the practical applications of the SASR to real DNA sequences should conform to this principle. Figure 4.2 shows the TP walk result of the 1st coding region (3,308 ~ 4,264) from the *Homo sapiens* (Human) mitochondrial DNA sequence (GenBank no. NC_001807) and Figure 4.3 is the TP walk of the sequence before this region (1 ~ 3,307, non-coding region without TP). It is easy to observe the trend in Figure 4.2, compared with the randomness in Figure 4.3. In Figure 4.2a, the walk moves rightward from (0, 0) to around (200, 0) in the complex plane within only 957 steps, but in Figure 4.3a, the walk moves around the zero point (real part: -50 ~ 0; imaginary part: -8 ~ 30) in the total 3,307 steps. Meanwhile, in Figure 4.2b, the real part keeps increasing with the growth of t while the imaginary part keeps relatively constant, and in Figure 4.3b, both the real part and the imaginary part oscillate without any pattern.

The simple example above shows that the TP walk visually discriminates simple coding sequences from non-coding ones. However, it is significant to quantitatively investigate, to what extent, the TP walk's patterns of simple coding and non-coding sequences are different, in order to check whether the coding regions can be further pointed out manually or computationally from the graph of a complete (coding/non-coding regions mixed) DNA sequence. Hence, the difference is analyzed between the walks of sequences in two typical datasets.

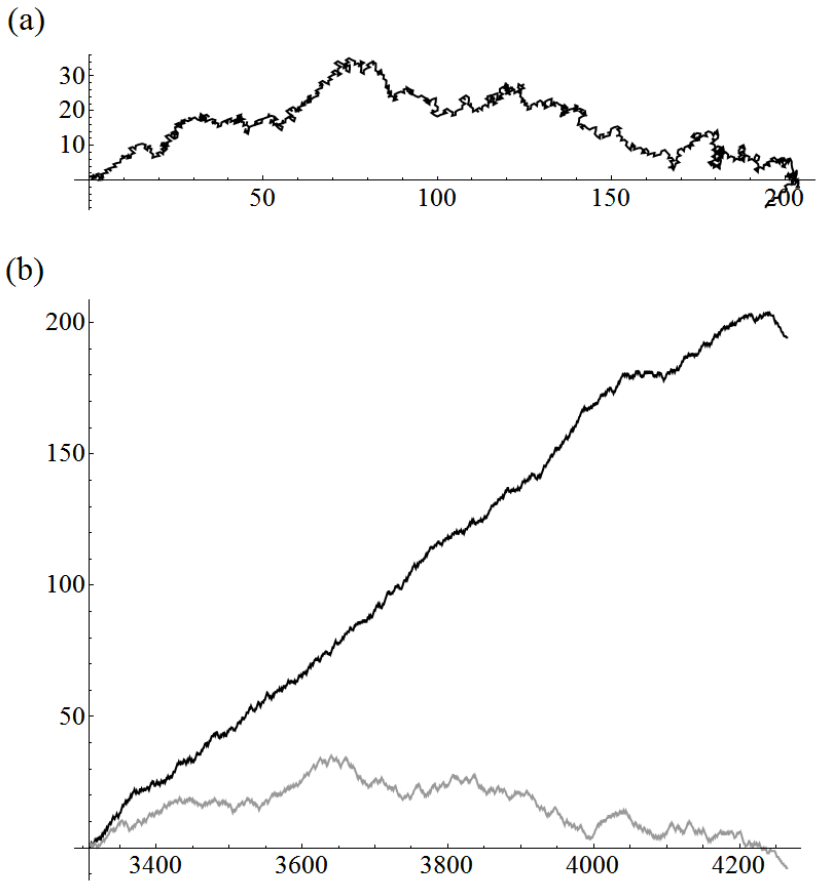


Figure 4.2 The TP walk of the 1st coding region (3,308 ~ 4,264) from the *Homo sapiens* (Human) mitochondrial DNA sequence (a) Walk trace in the complex plane (b) Plot of the real part (black) and the imaginary part (gray) of the points in the trace against the growing value of position *t*

The datasets are extracted from the first 15 chromosome DNA sequences of *S. cerevisiae* (GenBank no. NC_001133 ~ NC_001147). One is called here the coding set or the positive set, containing all of the single-exon genes with "experimental evidence". The other one is called the non-coding set or the negative set, containing all the inner sequences between genes.

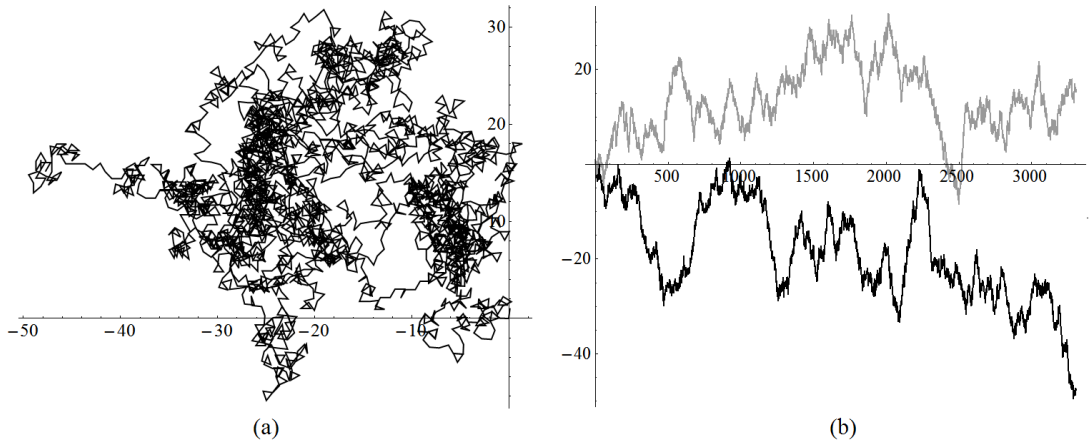


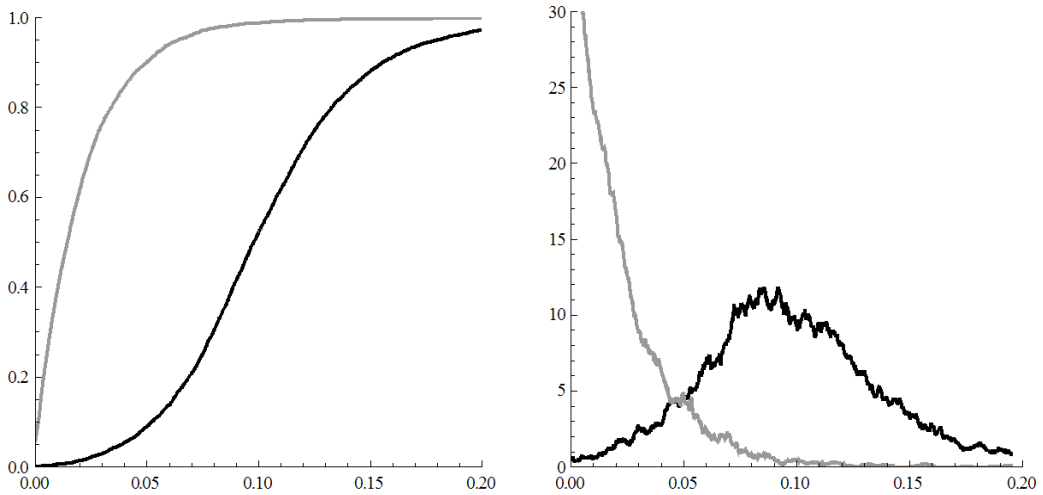
Figure 4.3 The TP walk of the sequence before the 1st coding region (1 ~ 3,307, non-coding region without the TP property) of the *Homo sapiens* (Human) mitochondrial DNA (a) Walk trace in the complex plane (b) Plot of the real part (black) and the imaginary part (gray) of the points in the trace against the growing value of position *t*

The SASR is applied to the sequences in these two datasets and the RR values are calculated (Section 3.3.1). The Cumulative Distribution Function (CDF) and the Probability Density Functions (PDF) of the RR distributions in the two datasets are plotted in Figure 4.4. As expected, the non-coding sequences occupy the low RR area and the coding sequences tend to be with higher RR values. The sample means *m* and the sample standard deviations *d* are listed in Table 4.1. An independent 2-sample *t*-test was conducted on these two distributions, in which the *t* value is found to be at 107.069 and the *p*-value at approximately 0. It indicates extremely high statistical significance of the difference between these two distributions.

OSCM (Anastassiou, 2000; Anastassiou, 2001) and SRM (Kotlar and Lavner, 2003) are also applied to the sequences in the two datasets. For OSCM, the four

coefficients have been set up, in Anastassiou's study (Anastassiou, 2000), as a = 0.1+0.12i, t = -0.3-0.2i, c = 0, and g = 0.45-0.19i, for the *S. cerevisiae* DNA. Meanwhile, the SRM is trained using the single-exon genes, which are with "experimental evidence" and in the forward reading direction, from the 16th chromosomes of *S. cerevisiae* (GenBank no. NC_001148). It is noticed that, the OSCM and the SRM, which are set up from the genes in the forward direction, may miss the TP property in the reverse coding sequences, which are also contained in the positive set. To recognize such a reverse TP property, in Anastassiou's study (2000) and Kotlar and Lavner's study (2003), complementary measures were involved. According to Anastassiou (2000), the four coefficients in the complementary measure are: $\tilde{a} = t'e^{-i2\pi/3}$, $\tilde{t} = a'e^{-i2\pi/3}$, $\tilde{c} = g'e^{-i2\pi/3}$, and $\tilde{g} = c'e^{-i2\pi/3}$. Here a', t', c', and g' are the complex conjugates of the original coefficients a, t, c, and g. In Kotlar and Lavner's study (2003), the complex coefficients of the four spectrums were also transformed in the same way to form the complementary measure. Therefore, the practical OSCM (or SRM) for discriminating coding (in both reading directions) and non-coding sequences is the greater one between the original measure and its complementary measure. The OSCM and the SRM are calculated for each entire sequence in the two datasets and the distributions of the measure values are obtained. The sample means and the sample standard deviations are also listed in Table 4.1. t-tests obtain p-values at 6.18×10^{-178} (OSCM) and 2.00×10^{-201} (SRM) for the

difference between the distributions of the positive and negative sets. Though they also show extremely high statistical significance of the difference, the *p*-values are higher and the *t* values are much less than the corresponding values obtained by using the RR measure. It reveals that more obvious difference is obtained between the two datasets by using the SASR than using the OCSM and the SRM.



(a) (b)

Figure 4.4 The RR distributions in the coding set (black) and the non-coding set (gray) (a) The Cumulative Distribution Function (CDF) (b) The Probability Density Function (PDF)

Table 4.1 Statistics of measures for the two DNA sequence datasets

		Size	RR		OSCM		SRM	
			m	*d*	*m*	*d*	*m*	*d*
Coding set (positive)		4144	0.10255	0.04444	0.00150	0.00099	0.04646	0.03289
Non-coding set (negative)		5594	0.02103	0.02401	0.00053	0.00221	0.01493	0.06558
2-sample *t*-test	*t* value		107.069		29.132		31.075	
	Degree of freedom		5924		8685		8207	
	p-value		0		6.18×10^{-178}		2.00×10^{-201}	

From another point of view, a classification using an RR threshold is investigated, in which a sequence is classified into coding if its RR value is beyond a threshold x, and non-coding otherwise. The sensitivity and the specificity of this classification are considered as follows (Zhang and Wang, 2000; Yin and Yau, 2007; Te Boekhorst et al., 2008):

$$\begin{aligned} \mathrm{Sn} &= \frac{\text{number of correctly classified coding sequences}}{\text{number of coding sequences}} \\ \mathrm{Sp} &= \frac{\text{number of correctly classified non-coding sequences}}{\text{number of non-coding sequences}} \end{aligned} \tag{4-1}$$

These two values, when using a given threshold x, can be easily derived from the CDF of the two RR distributions mentioned above: $\mathrm{Sn}(x) = 1 - F_p(x)$ and $\mathrm{Sp}(x) = F_n(x)$, where $F_p(x)$ and $F_n(x)$ are the CDF of the RR distributions in the coding set and the non-coding set respectively. The sensitivity and specificity are plotted in Figure 4.5a. It shows that both the sensitivity and specificity can reach about 90.5% at an RR threshold of about 0.05, over all the samples. Meanwhile, the OSCM and the SRM were also used, instead of the RR measure, for the same classification. The sensitivity and specificity are also derived from the CDF of the corresponding distributions. The averages of Sn and Sp by using these two measures are plotted in Figure 4.5b, together with the corresponding values obtained by using the RR measure. It shows that, the peaks can reach only 83.5% and 85% by using the OSCM and the SRM respectively, which shows less accuracy, compared with that obtained from the RR measure.

Besides, the sequences are cataloged by their lengths and the RR threshold is fixed at 0.05. It is found that the sensitivity in recognizing the long coding sequences is higher than that in recognizing the short ones (see Table 4.2), and the specificity shows a similar change over catalogs, except for a significant drop at the "longest length catalog", i.e., > 3,300bp. However, when the threshold is raised to 0.075, the coding regions in this catalog can be well discriminated with a Sn / Sp of 92.8% / 98.7%. It shows that the walk patterns of the very long (> 3,300bp) coding and non-coding sequences still differ enough for the discrimination, and the low specificity, when using the threshold of 0.05, may be caused by other periodicity patterns (unrelated to genetic coding). Beside, the precision is also calculated as follows (Olson and Delen, 2008):

$$\mathrm{Pr} = \frac{\text{number of correctly classified coding sequences}}{\text{number of classified coding sequences}} \tag{4-2}$$

This value reflects the reliability of the classified coding sequences, which is impacted from the capability of the classification method as well as the coding/non-coding proportion of the dataset. Table 4.2 shows that the precision is low in classifying sequences shorter than 600bp, compared with the high values in the long length catalogs, although the very biased coding/non-coding proportion of the dataset should also be noticed in the shortest and longest catalogs. In general, this test reveals that the TP walk is more capable of discriminating longer regions than shorter ones, despite in the view of sensitivity, specificity, or precision. This result is

reasonable, because this approach is essentially to visualize the persistency of the local TP and it statistically requires a sufficient length to show such a persistency.

Table 4.2 The sensitivity (Sn), specificity (Sp), and precision (Pr) in recognizing coding sequences with different lengths using the fixed RR threshold of 0.05. I: Number of the coding sequences; II: Number of the coding sequences classified as coding sequences; III: Number of the coding sequences classified as non-coding sequences; IV: Number of the non-coding sequences; V: Number of the non-coding sequences classified as non-coding sequences; VI: Number of the non-coding sequences classified as coding sequences. In the row with *, threshold 0.075 is used.

Length	I	II (Sn)	III	IV	V (Sp)	VI	Pr
1 ~ 300	89	66 (74.2%)	23	2176	1813 (83.3%)	363	15.4%
301 ~ 600	463	364 (78.6%)	99	1940	1602 (92.6%)	338	51.8%
601 ~ 900	612	524 (85.6%)	88	710	694 (97.7%)	16	97.0%
901 ~ 1200	656	581 (88.6%)	75	295	293 (99.3%)	2	99.7%
1201 ~ 1500	552	513 (92.9%)	39	156	156 (100%)	0	100%
1501 ~ 1800	493	472 (95.7%)	21	98	96 (98.0%)	2	99.6%
1801 ~ 2100	340	324 (95.3%)	16	59	57 (96.6%)	2	99.4%
2101 ~ 2400	232	227 (97.8%)	5	40	40 (100%)	0	100%
2401 ~ 2700	190	183 (96.3%)	7	29	29 (100%)	0	100%
2701 ~ 3000	122	121 (99.2%)	1	16	16 (100%)	0	100%
3001 ~ 3300	103	102 (99.0%)	1	18	18 (100%)	0	100%
3301 ~ ∞	292	291 (99.5%)	1	57	33 (57.9%)	24	92.4%
* 3301 ~ ∞	292	271 (92.8%)	21	57	56 (98.7%)	1	99.6%

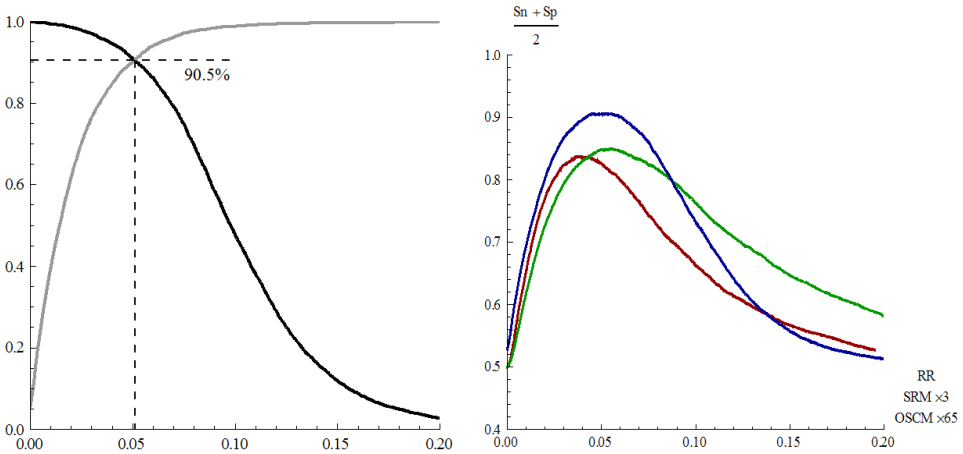


Figure 4.5 The accuracies in classifying sequences (a) The sensitivity (black) and specificity (gray) in the classification by using the RR measure (b) The averages of the sensitivity and specificity in the classification by using the OSCM (red), the SRM (green), and the RR measure (blue), respectively

The above experiments confirm the behaviors of TP walks of simple coding and non-coding sequences as described in Section 3.2.1.

4.2.2 TP walks of C_0-I-C_1 chains

In order to evaluate practical behaviors of TP walks of C_0-I-C_1 chains in real DNA sequences, 12 chains are randomly selected from a set of mitochondrial DNA sequences, in which there are 9 coherent (3 chains for each Δ = 0, 1, 2) and 3 incoherent. Table 4.3 shows more details about these 12 chains.

Figures 4.6 ~ 4.9 show the TP walk of the 12 chains after the application of the SASR. Figure 4.6 and Figure 4.7 are the visualization results, in which the coding

regions C_0 and C_1 are indicated by the obvious moving trends in the walks, and in the coherent cases, the corner shapes between trends follow the corner rule. Besides, it is clear from Figure 4.8 and Figure 4.9 that the walks keep relatively constant in the non-coding regions I compared with the high speed moving in the coding regions C_0 and C_1, regardless of whether the chain is coherent or not.

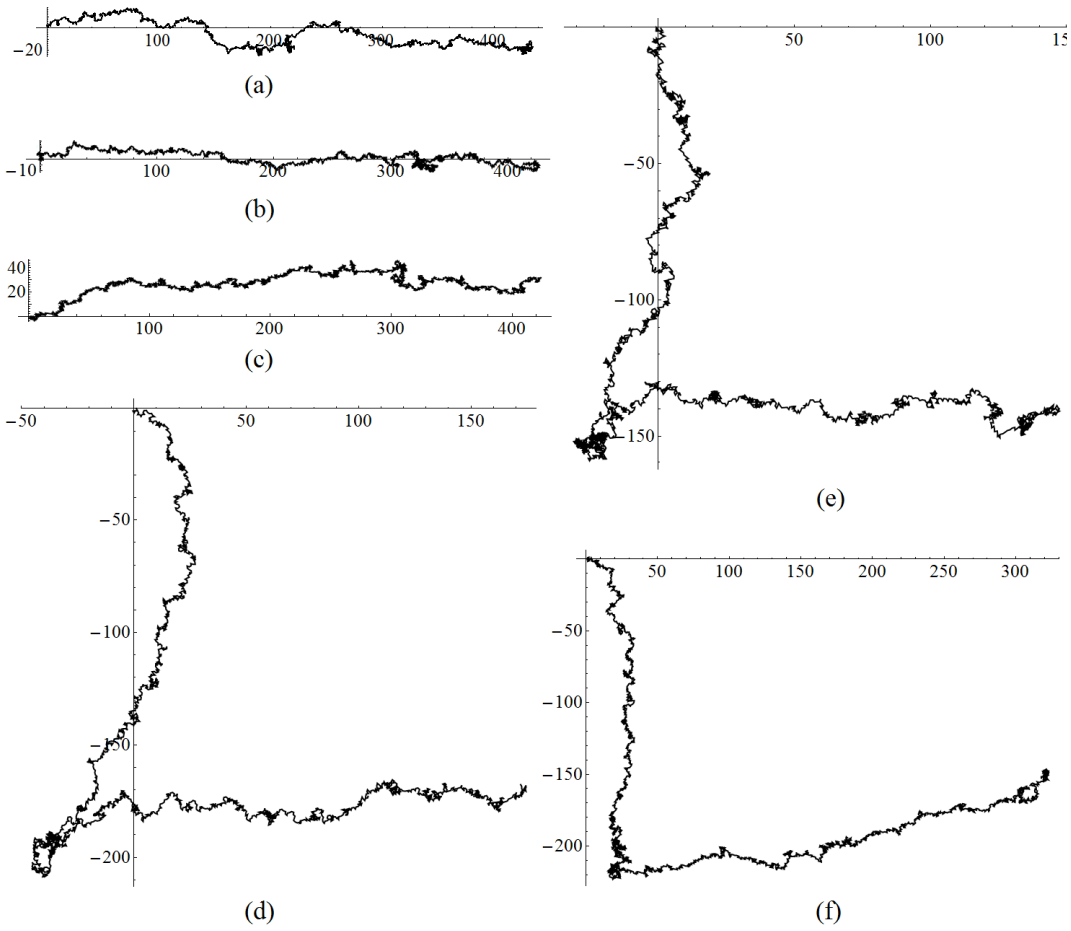


Figure 4.6 The TP walks of some C_0-I-C_1 chains in the complex plane (a ~ f) for chains 1 ~ 6 in Table 4.3

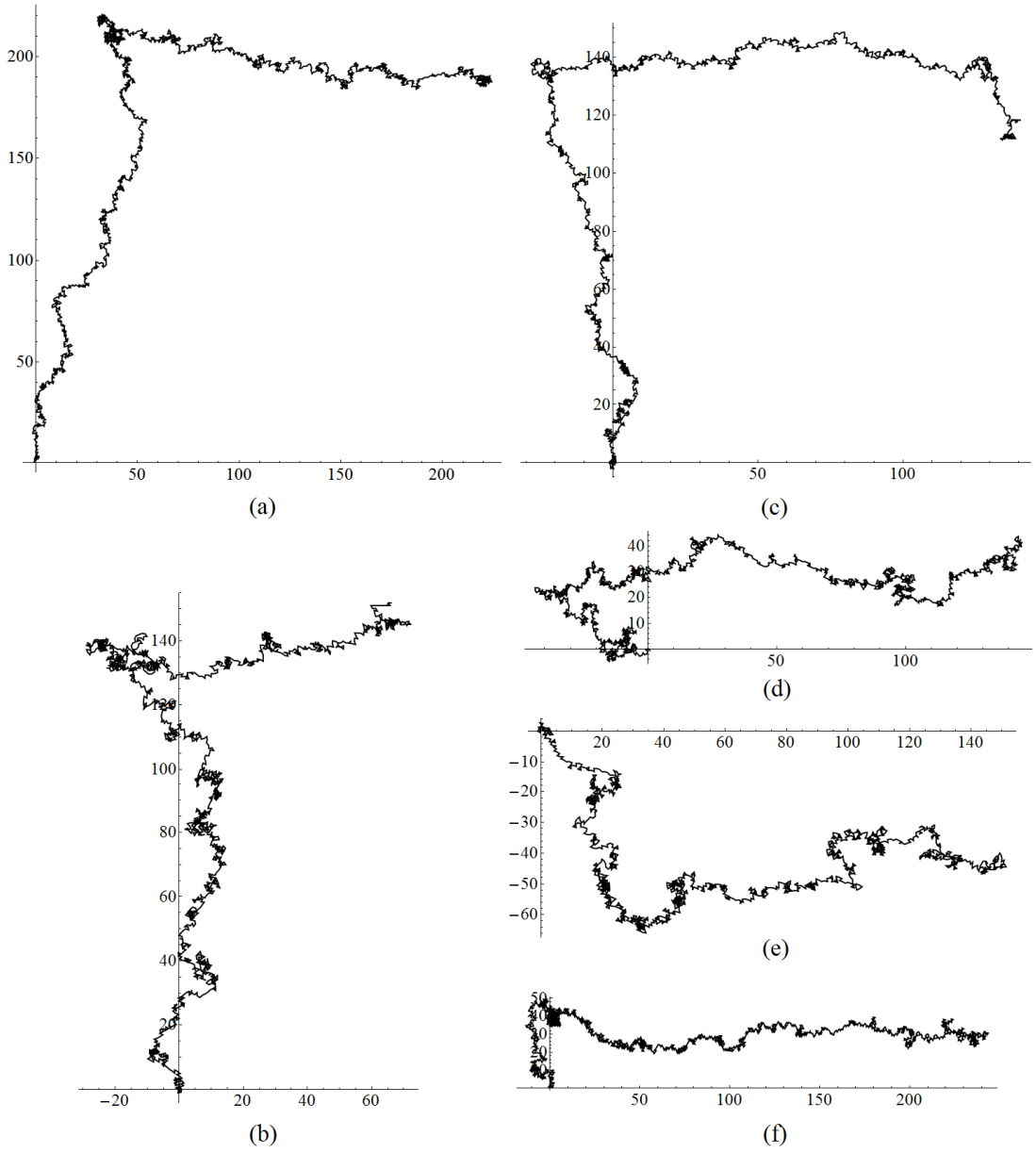


Figure 4.7 The TP walks of some C_0-I-C_1 chains in the complex plane (a ~ f) for chains 7 ~ 12 in Table 4.3

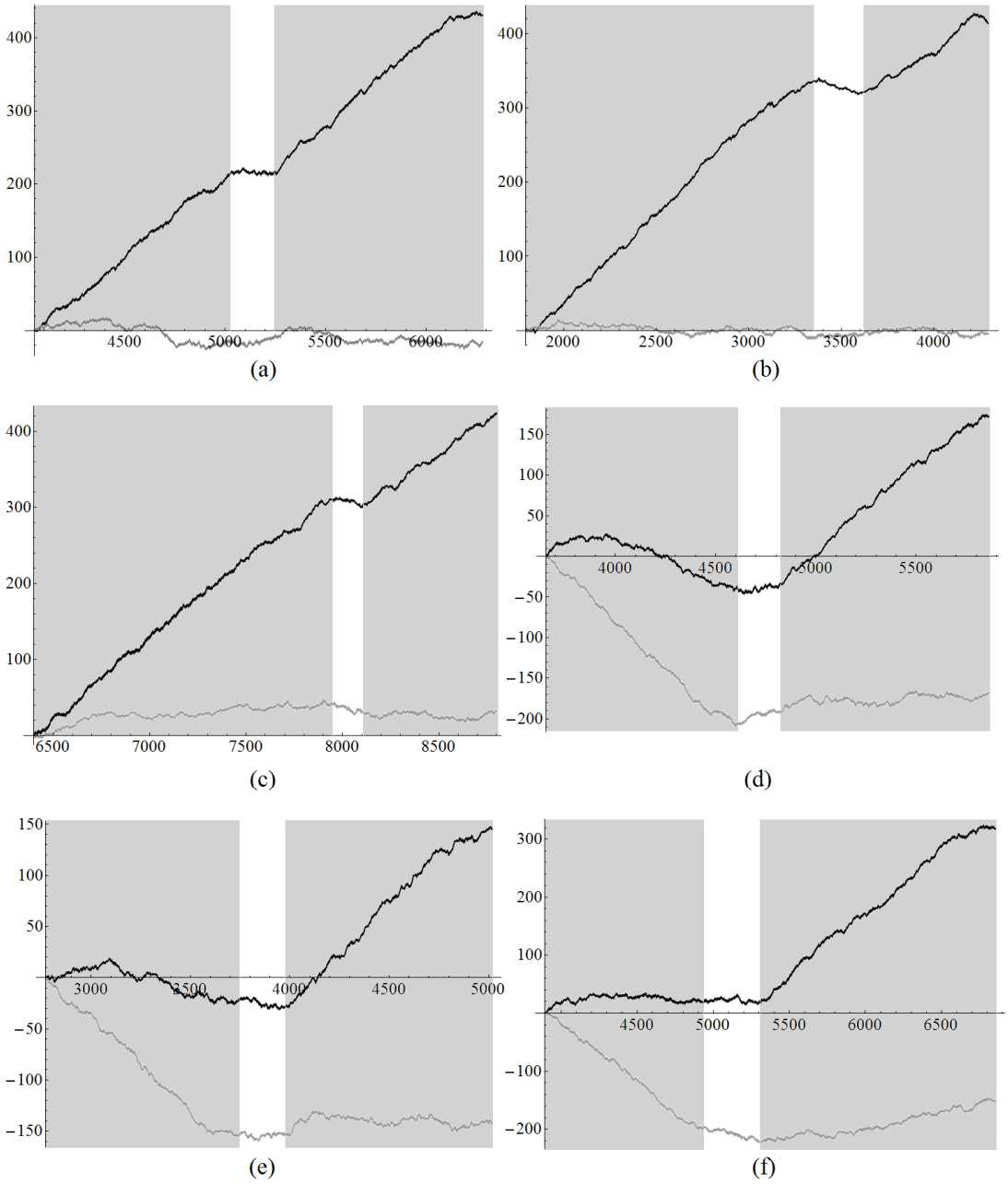


Figure 4.8 Plots of the real part (black) and the imaginary part (gray) against the sequence position *t* during the TP walks (a ~ f) for chains 1 ~ 6 in Table 4.3

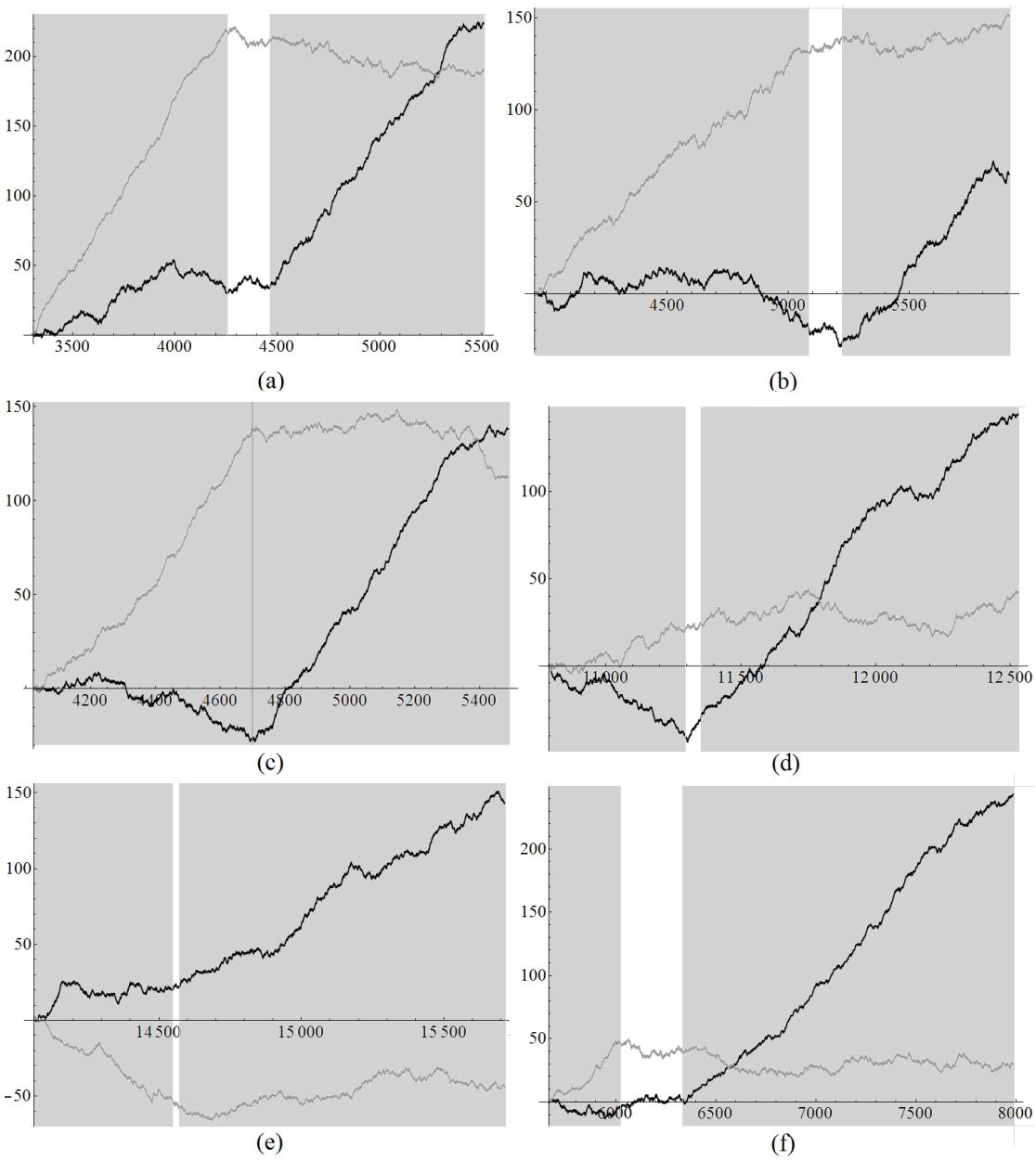


Figure 4.9 Plots of the real part (black) and the imaginary part (gray) against the sequence position t during the TP walks (a ~ f) for chains 7 ~ 12 in Table 4.3

These cases show that the walks of C_0-I-C_1 chains in real DNA sequences conform to the discussion in Section 3.2.2. Moreover, a qualitative verification of the corner rule will be presented in the next section.

Table 4.3 Some typical C_0-I-C_0 chains in real mitochondrial DNA sequences

Organism	Interval	C_0	C_1	Δ
1 *Gallus gallus* (NC_001323)	4,050 ~ 6,281	4,050 ~ 5,024	5,241 ~ 6,281	0
2 *Apis mellifera ligustica* (NC_001566)	1,794 ~ 4,295	1,794 ~ 3,357	3,618 ~ 4,295	0
3 *Cyprinus carpio* (NC_001606)	6,399 ~ 8,798	6,399 ~ 7,949	8,109 ~ 8,798	0
4 *Halichoerus grypus* (NC_001602)	3,654 ~ 5,862	3,654 ~ 4,610	4,819 ~ 5,862	1
5 *Rhea americana* (NC_000846)	2,771 ~ 5,018	2,771 ~ 3,745	3,978 ~ 5,018	1
6 *Rattus norvegicus* (AC_000022)	3,892 ~ 6,853	3,892 ~ 4,929	5,309 ~ 6,853	1
7 *Homo sapiens* (NC_001807)	3,308 ~ 5,511	3,308 ~ 4,264	4,471 ~ 5,511	2
8 *Lumbricus terrestris* (NC_001673)	3,952 ~ 5,912	3,952 ~ 5,088	5,220 ~ 5,912	2
9 *Anopheles gambiae* (NC_002084)	4,023 ~ 5,488	4,023 ~ 4,703	4,703 ~ 5,488	2
10 *Eptatretus burgeri* (NC_002807)	10,788 ~ 12,518	(–)10,788 ~ 11,291	11,361 ~ 12,518	–
11 *Arbacia lixula* (NC_001770)	14,062 ~ 15,713	(–)14,062 ~ 14,550	14,571 ~ 15,713	–
12 *Melipona bicolor* (NC_004529)	5,666 ~ 7,985	5,666 ~ 6,019	(–)6,339 ~ 7,985	–

4.2.3 TP walks of complete DNA sequences

In Section 3.2.3, an example is shown of the TP walk for the complete human mitochondrial DNA sequence. Besides, some more experiments are conducted on real DNA data, showing that the principle stands universally. Figures 4.10 ~ 4.12 are the TP walks of three complete mitochondrial DNA sequences with coding regions

marked in different colors. Table 4.4 shows details of these sequences. It is clear in the figure that the coding regions stay on arcs while the non-coding regions stay on corners or around stable points. In the (b) parts of the figures, the curves of the real part and the imaginary part fluctuate with the alternations of the coding and non-coding regions. Moreover, the principle of corner rule stands for every local coherent chain: the walk changes the direction on each corner according to the frame shift value Δ.

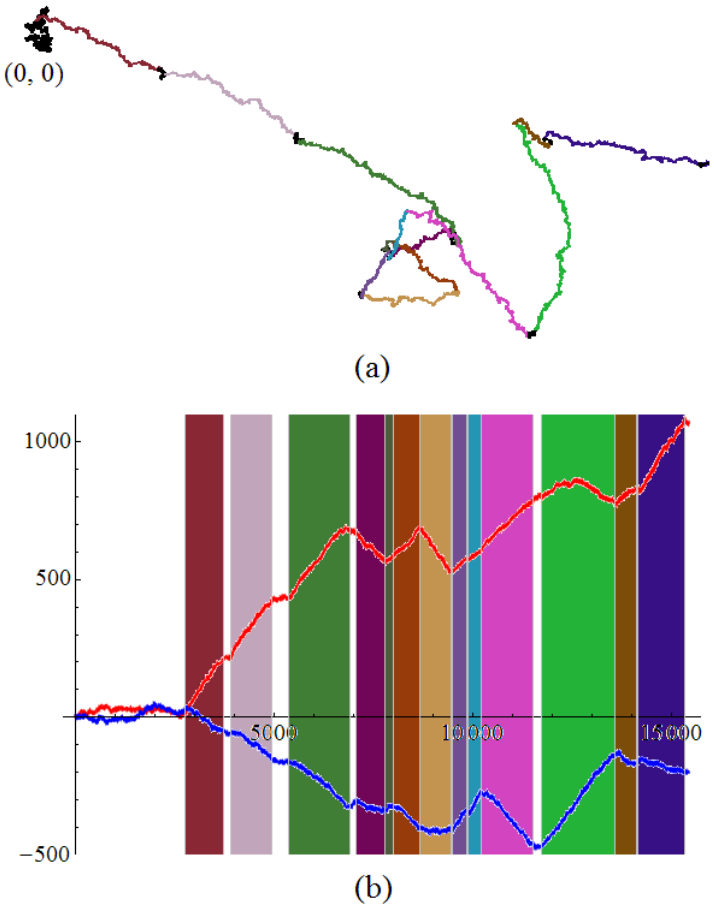


Figure 4.10 The TP walk of the complete mitochondrial DNA sequence from *Arctocephalus forsteri* (a) The walk in the complex plane (b) Plot of the real part (red) and the imaginary part (blue) of the walk against the position value *t*

Table 4.4 Coding regions in three complete mitochondrial DNA sequences

Arctocephalus forsteri (NC_004023)		*Emeus crassus* (NC_002673)		*Myxine glutinosa* (NC_002639)	
Interval	Δ	Interval	Δ	Interval	Δ
2,755 ~ 3,708		4,299 ~ 5,270		1 ~ 957	
3,922 ~ 4,962	0	5,489 ~ 6,529	2	1,180 ~ 2,226	0
5,356 ~ 6,900	0	6,890 ~ 8,440	0	2,564 ~ 4,117	1
7,041 ~ 7,724	2	8,577 ~ 9,260	1	4,247 ~ 4,936	0
7,796 ~ 7,999	2	9,333 ~ 9,500	0	4,995 ~ 5,159	1
7,957 ~ 8,637	2	9,491 ~ 10,174	2	5,153 ~ 5,839	2
8,637 ~ 9,419	2	10,174 ~ 10,956	2	5,842 ~ 6,624	2
9,491 ~ 9,835	2	11,027 ~ 11,199	1	6,714 ~ 7,061	2
9,908 ~ 10,204	0	Joined 〈 11,201 ~ 11,378	1	7,130 ~ 7,420	2
10,198 ~ 11,574	2	11,448 ~ 11,744	0	7,414 ~ 8,790	2
11,774 ~ 13,600	1	11,738 ~ 13,114	2	8,983 ~ 10,788	0
(–) 13,584 ~ 14,111	–	13,324 ~ 15,141	2	(–) 10,784 ~ 11,287	–
14,185 ~ 15,324	–	15,152 ~ 16,294	1	11,355 ~ 12,512	–
		(–) 16,473 ~ 16,994	–		

In order to qualitatively verify the corner rule in real complete DNA sequences, the SASR is applied to all 16 chromosome DNA sequences of *S. cerevisiae* (GenBank no. NC_001133 ~ NC_001148).

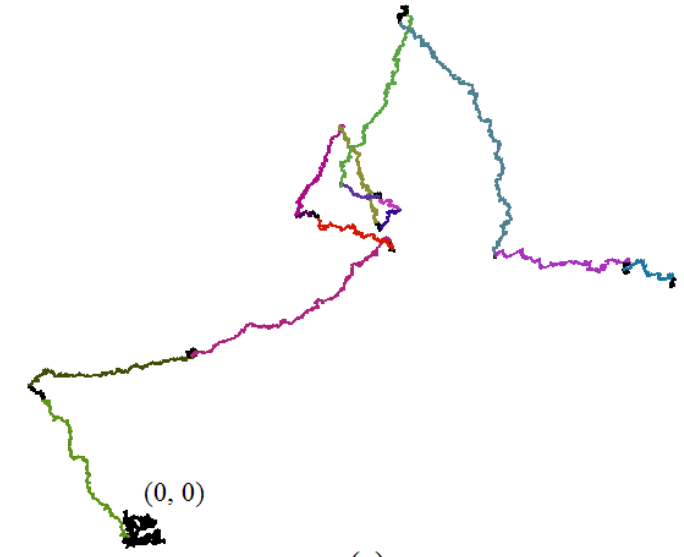
For each coherent local C_0-I-C_1 chain in the sequences, the direction shift of the TP walk on the corner between C_0 and C_1 is calculated as:

$$\alpha = \arg(z_{1e} - z_{1s}) - \arg(z_{0e} - z_{0s}) \tag{4-3}$$

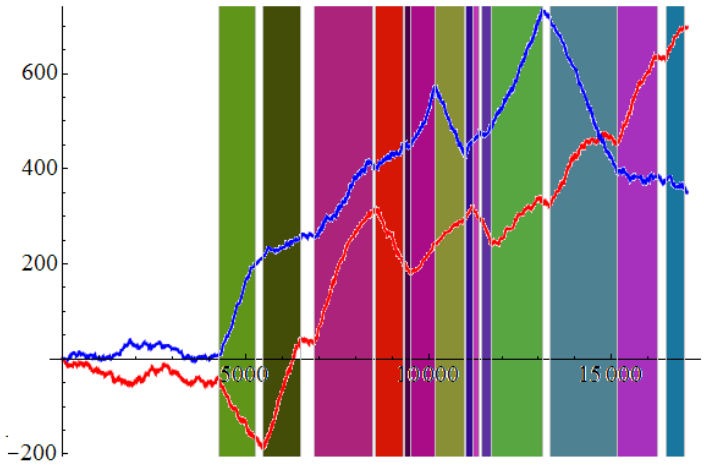
Here, the function arg(z) stands for the phase angle of a given complex number z. z_{0s}, z_{0e}, z_{1s}, and z_{1e} stand for the start point of C_0, the end point of C_0, the start point of C_1, and the end point of C_1, respectively. After that, α is further moved into the

domain $(-\pi, \pi]$ by $\pm 2\pi$:

$$\alpha \to \begin{cases} \alpha + 2\pi & \alpha \le -\pi \\ \alpha & -\pi < \alpha \le \pi \\ \alpha - 2\pi & \pi < \alpha \end{cases}$$



(a)



(b)

Figure 4.11 The TP walk of the complete mitochondrial DNA sequence from *Emeus crassus* (a) The walk in the complex plane (b) Plot of the real part (red) and the imaginary part (blue) of the walk against the position value *t*

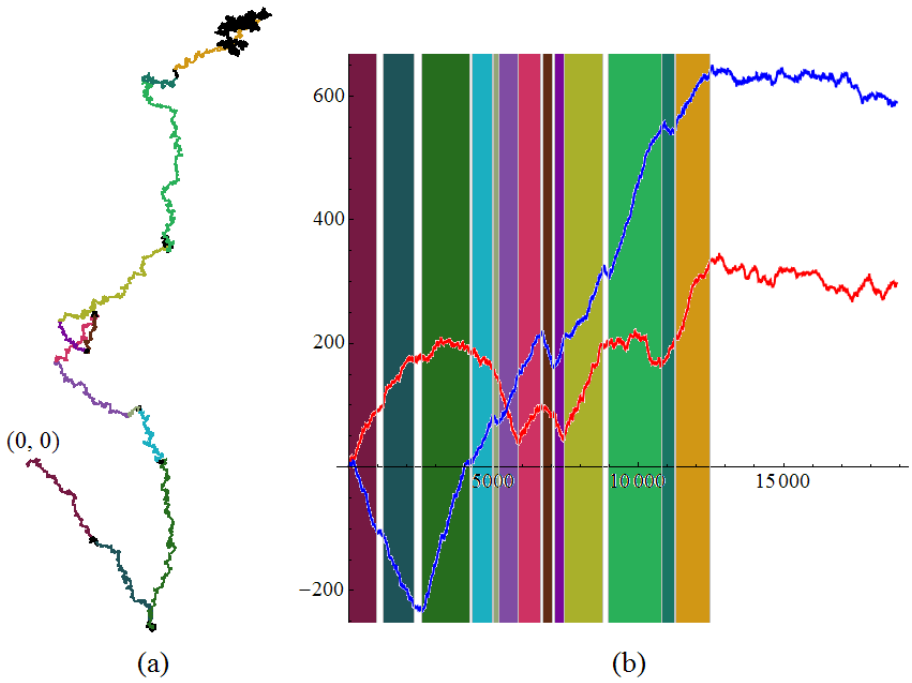


Figure 4.12 The TP walk of the complete mitochondrial DNA sequence from *Myxine glutinosa* (a) The walk in the complex plane (b) Plot of the real part (red) and the imaginary part (blue) of the walk against the position value *t*

For chains with each frame shift value $\Delta = 0$, 1, or 2, a distribution of α is obtained and plotted in Figure 4.13. It is clearly shown that: For $\Delta = 0$, α distributes mainly around 0, implying that the direction does not change and the walk moves straightly during those C_0-I-C_1 chains. For $\Delta = 1$ or 2, the distribution is found with an obvious high peak at $2\pi/3$ or $-2\pi/3$, respectively. It implies turning left or right on the corners of those chains. Table 4.6 shows some statistics in detail. The mean values of α are close to 0, $2\pi/3$, and $-2\pi/3$ for $\Delta = 0$, 1, and 2, respectively, and the sample standard deviations α_{dev} are low. In approximately 75% individual cases, α

are constrained in domains ($\alpha_{avg} - \alpha_{dev}$, $\alpha_{avg} + \alpha_{dev}$). It shows a high correlation between the frame shifts Δ and the direction shifts α in TP walks. These qualitative results are fully consistent with the discussion about the corner rule in Section 3.2.2.

Moreover, a similar experiment is conducted on a dataset from more comprehensive mitochondrial DNA sequences. 50 mtDNA sequences (listed in Table 4.5) are randomly selected and the direction shifts on the corners of the coherent local chains are investigated. The statistics are also presented in Table 4.6. The mean values are close to those from the chromosome DNA sequences and the sample standard deviations are much lower. In 92% individual cases, α are constrained in domains ($\alpha_{avg} - \alpha_{dev}$, $\alpha_{avg} + \alpha_{dev}$). It shows that the corners reveal the frame shifts for mtDNA sequences even better than those for chromosome DNA sequences.

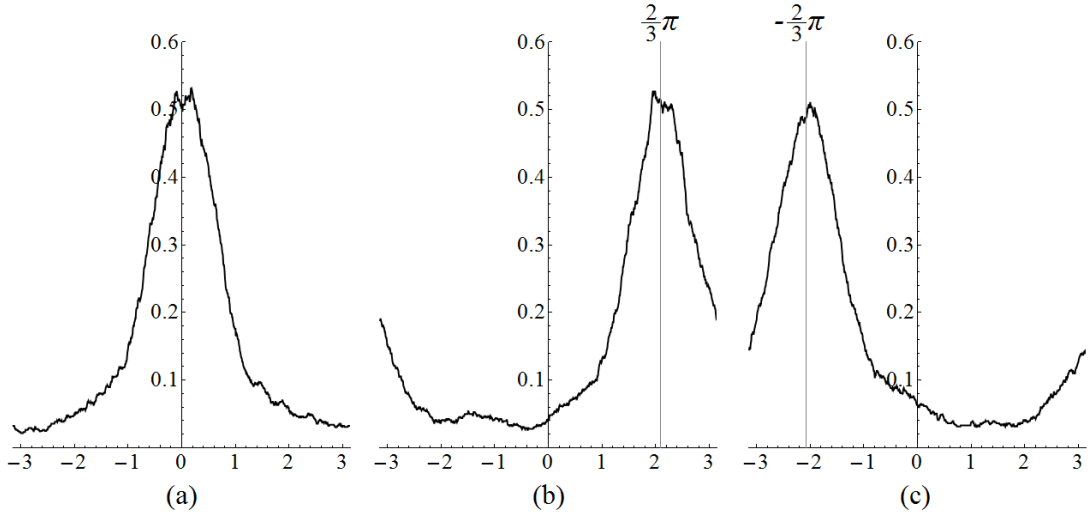


Figure 4.13 The distributions (in PDF) of the direction shifts α over C_0-I-C_1 chains (a) For chains with $\Delta = 0$ (b) For chains with $\Delta = 1$ (c) For chains with $\Delta = 2$

Table 4.5 List of the mitochondrial DNA sequences from 50 species

GenBank no.	Species	GenBank no.	Species
NC_003181	*Polymixia lowei*	NC_003196	*Pagrus major*
NC_004378	*Lamprogrammus niger*	NC_004394	*Ostichthys japonicus*
NC_004401	*Indostomus paradoxus*	NC_004622	*Antheraea pernyi*
NC_004696	*Lefua echigonia*	NC_004920	*Chrysochloris asiatica*
NC_005089	*Mus musculus*	NC_005275	*Platanista minor*
NC_005293	*Ixodes holocyclus*	NC_005313	*Auxis rochei*
NC_005779	*Drosophila mauritiana*	NC_005796	*Pterothrissus gissu*
NC_005932	*Ninox novaeseelandiae*	NC_005934	*Armillifer armillatus*
NC_006131	*Acanthogobius hasta*	NC_006160	*Aleurochiton aceris*
NC_006283	*Diplometopon zarudnyi*	NC_006329	*Eurycea bislineata*
NC_006335	*Plethodon elongatus*	NC_006405	*Kaloula pulchra*
NC_006408	*Polypedates megacephalus*	NC_006533	*Anguilla australis*
NC_006538	*Anguilla dieffenbachii*	NC_006839	*Xenopus (Silurana) tropicalis*
NC_006886	*Mytilus galloprovincialis*	NC_006890	*Ambystoma californiense*
NC_006919	*Sundasalanx mekongensis*	NC_006925	*Mystacina tuberculata*
NC_007175	*Crassostrea virginica*	NC_007215	*Lepeophtheirus salmonis*
NC_007231	*Oreochromis mossambicus*	NC_007240	*Gallus sonneratii*
NC_007402	*Xenopeltis unicolor*	NC_007442	*Gonodactylus chiragra*
NC_007630	*Dasyurus hallucatus*	NC_007693	*Manouria emys*
NC_007699	*Testudo kleinmanni*	NC_007789	*Acanthaster brevispinus*
NC_007883	*Menura novaehollandiae*	NC_007936	*Cricetulus griseus*
NC_007976	*Coreana raphaelis*	NC_008081	*Pseudohynobius tsinpaensis*
NC_008085	*Batrachuperus tibetanus*	NC_008089	*Onychodactylus fischeri*
NC_008109	*Scomberomorus cavalla*	NC_008125	*Lophius litulon*
NC_008132	*Nipponia nippon*	NC_008327	*Solea senegalensis*

Table 4.6 Statistics of the direction shifts α in TP walks, for coherent local C_0-I-C_1 chains with $\Delta = 0$, 1, and 2

	$\Delta = 0$	$\Delta = 1$	$\Delta = 2$
From chromosome DNA sequences of *S. cerevisiae*			
α_{avg} (mean value)	0.05	2.15	-2.04
α_{dev} (sample standard deviation)	1.02	1.07	1.08
P_0 (proportion in $\alpha_{avg} \pm \alpha_{dev}$)	75%	75%	74%
From 50 mitochondrial DNA sequences			
α_{avg} (mean value)	0.02	2.04	-2.00
α_{dev} (sample standard deviation)	0.64	0.62	0.51
P_0 (proportion in $\alpha_{avg} \pm \alpha_{dev}$)	92%	92%	92%

4.3 Computational prediction of coding regions based on the SASR

In this section, experiments are conducted to test the performances of the computational methods, including the FSND and the T-Z-T analysis, in predicting the coding regions of real DNA sequences.

4.3.1 Applications of the FSND

In this section, the applications of the FSND are with a fixed analysis scale Δt = 127 as mentioned in Section 3.3.2. 12 mtDNA sequences are selected for the evaluation of the FSND method. 6 of them are used for training in order to obtain a regression function between the average moving speed v_{avg} and the optimal speed threshold v_{opt}, at which the highest accuracy can be achieved. After application of the SASR to these 6 sequences, v_{avg} and v_{opt} of each sequence are obtained in Table 4.7.

It shows that, at v_{opt}, the FSND may achieve accuracies of around 90% for this training set and a linear relationship is found between v_{avg} and v_{opt}. Therefore, a linear regression model is built to estimate v_{opt} from a given value of v_{avg}:

$$v_{opt} \approx h(v_{avg}) = 0.771 v_{avg} - 0.009$$

Table 4.7 Details in the application of the SASR-FSND to 6 mtDNA sequences in the training set

Species	GenBank no.	v_{avg}	v_{opt}	Accuracy ($v_0 = v_{opt}$)
Ursus maritimus	NC_003428	0.115	0.080	89%
Homo sapiens	NC_001807	0.154	0.117	90%
Rhea americana	NC_000846	0.142	0.090	94%
Gallus gallus	NC_001323	0.157	0.103	92%
Dinodon semicarinatus	NC_001945	0.152	0.116	90%
Apis mellifera ligustica	NC_001566	0.128	0.091	89%

After that, the SASR is applied to the rest 6 sequences in the dataset and v_{opt} is estimated for each sequence using the above linear regression model. As mentioned in Section 3.3.2, each position in the sequences is assigned to coding if the local moving speed of the walk is beyond an estimated value of v_{opt}. The performances of these predictions are presented in Table 4.8. The accuracies are also around 90%, which reveal similar performances to those achieved in applications to the training set. Therefore, the above linear regression model is said to be appropriate for the estimation of v_{opt}.

Table 4.8 Details in the application of the SASR-FSND to the rest 6 mtDNA sequences, after training with the 6 previous sequences

Species	GenBank no.	v_{avg}	v_{opt}	Accuracy ($v_0 = v_{opt}$)
Sus scrofa	NC_000845	0.154	0.110	93%
Rattus norvegicus	AC_000022	0.151	0.107	91%
Dasypus novemcinctus	NC_001821	0.162	0.116	90%
Platynereis dumerilii	NC_000931	0.131	0.092	90%
Bombyx mori	NC_002355	0.125	0.087	90%
Chauliodus sloani	NC_003159	0.120	0.087	87%

Furthermore, 2 sequences are removed from the training set, i.e., *Dinodon semicarinatus* mtDNA (GenBank no. NC_001945) and *Apis mellifera ligustica* mtDNA (GenBank no. NC_001566). After that, a different linear regression model is obtained:

$$v_{opt} \approx h(v_{avg}) = 0.718 v_{avg} - 0.004$$

Based on this new regression model, there is a newly estimated value of v_{opt} for each sequence. However, as shown in Table 4.9, the performances of coding predictions for the previous 6 sequences remain the same. This regression model is also used to estimate v_{opt} for the 2 sequences removed from the training set and the predictions for these 2 sequences achieve the optimal accuracies, which have already been tested previously. It implies that the training process of the FSND is not critical to the training set, a small relative set is sufficient to obtain an appropriate regression

model, and the change of the training set will not considerably influence the performance of the prediction.

Table 4.9 Details in the application of the SASR-FSND to the rest 8 mtDNA sequences, after training with 4 sequences

Species	GenBank no.	v_{avg}	v_{opt}	Accuracy ($v_0 = v_{opt}$)
Sus scrofa	NC_000845	0.154	0.106	93%
Rattus norvegicus	AC_000022	0.151	0.104	91%
Dasypus novemcinctus	NC_001821	0.162	0.112	90%
Platynereis dumerilii	NC_000931	0.131	0.090	90%
Bombyx mori	NC_002355	0.125	0.086	90%
Chauliodus sloani	NC_003159	0.120	0.085	87%
Dinodon semicarinatus	NC_001945	0.152	0.105	90%
Apis mellifera ligustica	NC_001566	0.128	0.088	89%

4.3.2 Applications of the T-Z-T analysis

The FSND assigns coding / non-coding to each site in a sequence, with a satisfactory accuracy as shown in the above section. However, as mentioned in Section 3.3.3, the sites are not organized in the form of regions, but considered individually. And the local moving speed is the only concern in the FSND. It causes the FSND to take less advantage of the "corner rule" in a TP walk to understand the frame shifts between coding regions. And moreover, a fixed analysis scale is involved, and a training process is still necessary, though it is not very critical. This

section evaluates the performance of the T-Z-T analysis, which segments a complete TP walk and recognizes coding region candidates without any training or fixed analysis scale, as presented in Section 3.3.3.

The significance levels P_0 and P_2 are set to 95% and 99% respectively, as recommended in Section 3.3.3. In order to get more opportunities to provide potential coding region candidates (with a high expected sensitivity), a relatively low specificity is expected as a tradeoff and here P_1 is set to 85%.

The T-Z-T analysis is applied to the TP walks of the 12 mtDNA sequences previously used in Section 4.3.1. Sensitivity (Sn), specificity (Sp), and accuracy (Ac) of the predictions are shown in Table 4.10. It shows that, without any preceding training process, the T-Z-T analysis achieves higher accuracies (Ac) ranging from 90% to 95%, compared with the results in Section 4.3.1 (see Table 4.7 ~ 4.9). In most of the cases, a great sensitivity (Sn) close to 100% is obtained and it shows a high capability of this approach in providing coding region "candidates" before any extra information is available. Compared with the high sensitivity, the specificity (Sp) is relatively lower, mainly because the expected specificity P_1 is set only 85%, in the configuration, as a tradeoff to obtain a high sensitivity. And it is also due to the lower sensitivity of this approach in detecting very short non-coding regions, which tend to mingle with the neighboring coding region candidates in the prediction.

Table 4.10 Performances of the T-Z-T analysis for 12 mtDNA sequences

Species	GenBank no.	Sn	Sp	Ac
Ursus maritimus	NC_003428	99.9%	76.6%	92.1%
Homo sapiens	NC_001807	99.6%	83.6%	94.5%
Rhea americana	NC_000846	96.6%	85.5%	93.0%
Gallus gallus	NC_001323	100%	74.5%	91.8%
Dinodon semicarinatus	NC_001945	96.9%	90.0%	94.5%
Apis mellifera ligustica	NC_001566	99.2%	70.2%	89.8%
Sus scrofa	NC_000845	97.5%	86.6%	94.1%
Rattus norvegicus	AC_000022	99.5%	85.9%	95.3%
Dasypus novemcinctus	NC_001821	99.0%	74.9%	90.9%
Platynereis dumerilii	NC_000931	98.7%	71.3%	90.6%
Bombyx mori	NC_002355	93.7%	80.0%	89.8%
Chauliodus sloani	NC_003159	95.8%	87.6%	92.8%

The prediction for the human (*Homo sapiens*) mtDNA sequence (GenBank no. NC_001807) is plotted in Figure 4.14. The blue bars in the figures indicate the predicted coding region candidates and the orange bars show the true regions. It shows that all the coding regions are predicted with the boundaries (marked in red) very close to the true boundaries. However, some short non-coding regions, e.g. the non-coding region between the 1st and the 2nd coding regions, mingle with the neighboring coding region candidates. Some local C_0-I-C_1 chains with very short non-coding regions I and $\Delta = 0$ (no frame shift), e.g. the chain containing the 4th and the 5th coding regions and the chain containing the 7th, the 8th, and the 9th coding regions, merge into long coding region candidates, which call for a better method and extra information for identification.

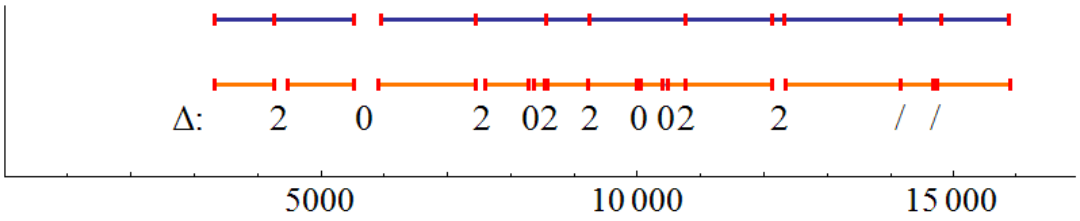


Figure 4.14 Plot of the coding region candidates provided by the T-Z-T analysis for the *Homo sapiens* mtDNA sequence (GanBank no. NC_001807). The blue bars indicate the predicted coding region candidates, the orange bars indicate the true regions from the experimental results, and the red vertical bars are the boundaries of the regions. The numbers below the orange bars show the frame shift value Δ between each two neighboring coding regions.

To investigate the advantages of the proposed method compared with popular gene prediction methods, some of the 12 mtDNA sequences in Table 4.10 are further analyzed with GeneMark.hmm (http://exon.biology.gatech.edu/) and GENSCAN (http://genes.mit.edu/GENSCAN.html). These two widely used online gene predicting tools are based on HMM algorithms. For the gene predictions in different target organisms, each of these two tools provides three models trained by different training datasets: GeneMark.hmm provides models for prokaryotes, low eukaryotes model and eukaryotes, while users of GENSCAN can choose the vertebrate model, Arabidopsis model or Maize model. After the mtDNA sequences are analyzed with these two tools using each of the models mentioned above, a set of predictions is obtained, and the performances of two of them are listed in Table 4.11. It is noticed

that, with very poor sensitivity, all the three models of GeneMark.hmm recognizes very few coding sites and results in low accuracies of less than 50%. The results from GENSCAN are even worse. The poor results from GeneMark.hmm and GENSCAN are expected, because the models are not well trained yet by suitable training datasets, e.g. an mtDNA training set, as they are supposed to. However, it reveals that the accuracies of such HMM-based gene predictions depend highly on the training models: It is undeniable that the accuracies can be extremely high if the models are well trained (see another experiment on *S. pombe* below). However, when there is insufficient information for the training process, like in this experiment, such methods may fail to recognize coding sites. It therefore limits the applications of the methods to unknown DNA sequences with new coding patterns. In contrast, Table 4.11 shows that the accuracies of the T-Z-T analysis for the two sequences are relatively high as they exceed 94%. Although the HMM-based methods may perform better if they are well trained, the comparison in Table 4.11 demonstrates that when there is insufficient information for the training process, the T-Z-T analysis achieves great results, which are much better than those obtained from the HMM-based methods. The self-adaptive feature of SASR allows for the revelation of new and unknown coding patterns and facilitates coding region prediction in DNA sequences without available training sets.

Table 4.11 Performances of the T-Z-T analysis for 2 mtDNA sequences, compared with those of GeneMark.hmm and GENSCAN

		Homo sapiens mtDNA			*Sus scrofa* mtDNA		
		Sn	**Sp**	**Ac**	**Sn**	**Sp**	**Ac**
T-Z-T analysis		99.6%	83.6%	94.5%	97.5%	86.6%	94.1%
GeneMark.hmm	**Prokaryotes model**	34.6%	98.6%	54.8%	27.9%	94.0%	48.8%
	Low eukaryotes model	28.0%	94.0%	48.9%	25.8%	95.0%	47.8%
	Eukaryotes model	0%	99.0%	31.3%	0.3%	99.5%	31.7%
GENSCAN	**Vertebrate model**	7.5%	100%	36.7%	7.3%	97.6%	36.0%
	Arabidopsis model	6.8%	94.6%	34.5%	19.5%	95.7%	43.7%
	Maize model	2.2%	100%	33.1%	1.0%	100%	32.4%

Another experiment is conducted on a fragment from the 2nd chromosome DNA sequence of *S. pombe* (GenBank no. NC_003423) in order to investigate the tolerance of these methods to input errors.

The T-Z-T analysis is applied to the TP walk of the DNA fragment ranging from position 110,821 to 113,520, which contains gene SPBC359.03c from position 111,309 to 113,048 (reverse reading order). The method detects a coding region candidate from position 111,315 to 113,042, which is very close to the true region. Besides, a short false candidate is predicted from position 113,043 to 113,195, but it can be eliminated when the expected specificity P_1 is increased to 95%. Meanwhile this DNA fragment is also scanned by GeneMark.hmm and GENSCAN with their built-in models trained with suitable training set for *S. pombe*. The results are plotted in Figure 4.15a. The predictions made by these two popular HMM-based methods

are exactly the true coding region. It shows that the HMM-based methods can work perfectly after the models are well trained with a suitable training set.

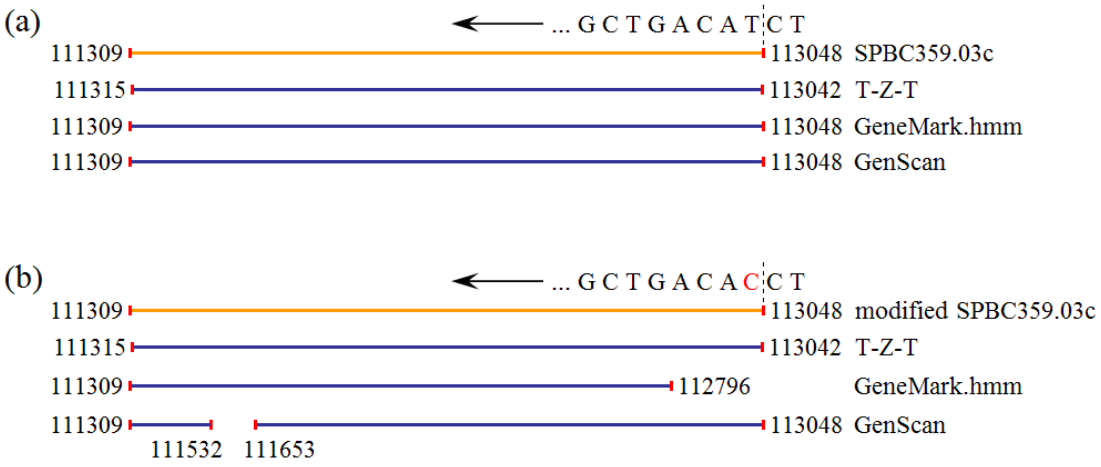


Figure 4.15 Plot of the prediction of SPBC359.03c in the 2nd chromosome DNA sequence of *S. pombe* by using the T-Z-T analysis, GeneMark.hmm, and GENSCAN. The blue bars indicate the predicted coding region candidates, the orange bars show the true regions from experimental results, and the red vertical bars are the boundaries of the regions. (a) Predictions of the original SPBC359.03c (b) Predictions of the modified SPBC359.03c with the "T" at position 113,048 changed into "C"

However, when some input errors are introduced into the DNA data, the results may be quite different. As a simulated input error, the base "T" at the coding start position (position 113,048) of gene SPBC359.03c is modified into "C". The T-Z-T analysis, GeneMark.hmm, and GENSCAN are applied to this modified sequence and the results are quite different, as shown in Figure 4.15b. It shows that an error at a

single position in the input sequence causes the predicted coding region to be shortened by 252 sites with GeneMark.hmm and the predicted region is broken down into two exons with GENSCAN. On the other hand, the predicted region from the T-Z-T analysis remains the same. For more comprehensive understanding of the influence of the rough input data, where errors may take place at random positions, two sets of modified sequences are further built from this original fragment. In one set, called “head set” (H set), each of the 33 sequences contains one modified site in the region ranging from position 113,043 (-5 from the coding start position) to 113,053 (+5 from the coding start position). And in the other set called “middle set” (M set), modifications are made in the middle of the coding region, from position 111,995 to 112,005. The three methods are applied to all 66 sequences in these two sets and the changes in the predictions after the modifications of the input data are investigated by calculating the number of sites with changed coding/non-coding assignments. Table 4.12 shows that the modifications near the coding start position (H set) cause great changes in the predictions from the HMM-based methods: On average over the 33 sequences in the H set, GeneMark.hmm has 84.0 sites that changed their assignments and GENSCAN has 44.1. On the other hand, the modifications in this region do not considerably influence the predictions made by the T-Z-T analysis: The average number of the changed assignments is only 2.6. Table 4.12 also indicates that the modifications in the middle of the coding region

give less influence to the predictions from all the three methods: When the methods are applied to the M set, the average number of changed assignments reduces to 2.8 for GeneMark.hmm and 20.8 for GENSCAN. And for the T-Z-T analysis, all 33 modified sequences in the M set produce the same predictions as the original one.

Table 4.12 The average number of the changed coding/non-coding assignments after the modifications of the original fragment

Sequence set	Number of sequences	Average number of the changed assignments (per sequence)		
		GeneMark.hmm	GENSCAN	T-Z-T analysis
Head (H) set	33	84.0	44.1	2.6
Middle (M) set	33	2.8	20.8	0

The applications of the methods to the modified sequences indicate that the performances of the HMM-based methods highly depend not only on training sets, but also on the quality of input data. Because every single site (input signal) is critical in determining the state in the Markov model, the predictions provided by these methods are accurate, but less robust: An input error at a single site, especially near the coding start position, may change the prediction to a great extent. On the other hand, input errors do not largely influence the SASR-based methods, e.g. the T-Z-T analysis, since the SASR deals with coding pattern recognition by considering the statistical property, i.e., the TP, in local groups of sites, rather than every single site. So, compared with the HMM-based methods, the SASR-based methods are more robust but less accurate and this feature is suitable for the early stage coding

prediction, when the input data is inaccurate.

4.4 Summary

Verifications of the newly proposed SASR approach have been presented in this chapter. A test of the computational time shows that generating a TP sequence from a DNA sequence is with a low computational complexity $O(N)$. The practical behaviors of TP walks have been verified using different kinds of real DNA sequences, including simple coding/non-coding sequences, C_0-I-C_1 chains, and complete DNA sequences. The computational predictions of protein-coding regions have been provided for real DNA sequences, by using the FSND and the T-Z-T after the SASR (denoted as SASR-FSND and SASR-TZT respectively). Among them, the SASR-TZT approach shows better performances. Compared with some HMM-based methods, the SASR-TZT approach requires no training, and meanwhile, it is more robust since a small input error does not greatly influence the results of the SASR-TZT. So it is viewed as a satisfactory solution to coding region prediction, especially for the early stage study on a newly sequenced DNA.

CHAPTER 5

EXTENSIONS OF THE SASR METHOD

In the above chapters, the application of the SASR to visualize coding regions in DNA sequences has been presented together with some further numerical analyses. The SASR is actually a tool for the investigation of periodicity-related phenomena in sequences and this method can be extended for applications in some other fields.

5.1 Study of nucleosome formation by a τ-periodicity SASR

In this section, the original SASR is extended for a τ-periodicity (τ is any rational number) and for dinucleotide sequences. The τ-period SASR for dinucleotide sequences is then adopted to investigate the relationship between the ~10bp periodicity and nucleosome formation.

5.1.1 τ-periodicity SASR

As mentioned before, the Fourier Transform on Voss's binary sequences gives:

$$U_{\Lambda}(k) = \sum_{t=1}^{N} u_{\Lambda}(t) e^{-i\frac{2\pi}{N}tk}$$

Consider the frequency $k = N/\tau$, where τ is a rational number that can be expressed by the quotient of two integers, i.e., $\tau = a/b$ and $k = bN/a$. Therefore the spectrum is:

$$U_\Lambda(\frac{bN}{a}) = \sum_{t=1}^{N} u_\Lambda(t) e^{-i\frac{2\pi b}{a}t} = \sum_{r=0}^{a-1}\left(e^{-i\frac{2\pi br}{a}} \sum_{t \bmod a = r} u_\Lambda(t) \right) = \sum_{r=0}^{a-1}\left(e^{-i\frac{2\pi (br\%a)}{a}} \cdot F_{\Lambda r} \right) \quad (5\text{-}1)$$

$$\text{denote: } F_{\Lambda r} = \sum_{t \bmod a = r} u_\Lambda(t) = \text{count}\{t \mid x_t = \Lambda \text{ and } t \bmod a = r\}$$

Similar to TPM, a "τ-Periodicity Matrix" is defined, denoted as τ-PM. A τ-PM is a 4×a matrix, where each row i stands for a nucleotide base (A, T, C, or G). And for each row $i = \Lambda$, the entry $m_{\Lambda j}$ (j = 1, 2, …, a) is defined as:

$$m_{\Lambda j} = \sum_{br\%a = j} F_{\Lambda r} \quad (5\text{-}2)$$

And the row vector $M_\Lambda = \{m_{\Lambda 1}, m_{\Lambda 2}, \ldots, m_{\Lambda a}\}$ is called a τ-periodicity vector. According to Equation (5-1) and Equation (5-2), we have:

$$U_\Lambda(\frac{N}{\tau}) = U_\Lambda(\frac{bN}{a}) = \sum_{j=1}^{a}\left(e^{-i\frac{2j\pi}{a}} \cdot m_{\Lambda j} \right) \quad (5\text{-}3)$$

Therefore, a mapping is built from a τ-periodicity vector to a Fourier spectrum. And shift operations on a τ-periodicity vector are equivalent to rotations on the corresponding complex number U:

$$M << rb \quad \leftrightarrow \quad U \cdot e^{i\frac{2rb\pi}{a}} = U \cdot e^{i\frac{2r\pi}{\tau}} \quad (\frac{2r\pi}{\tau} \text{ conter-clockwise rotation on } U)$$

$$M >> rb \quad \leftrightarrow \quad U \cdot e^{-i\frac{2rb\pi}{a}} = U \cdot e^{-i\frac{2r\pi}{\tau}} \quad (\frac{2r\pi}{\tau} \text{ clockwise rotation on } U)$$

It is easy to find that, TPM, TP vector, and the mapping from TP vectors to Fourier spectrums are just in a particular case of the τ-periodicity, where a = 3, b = 1, and $\tau = a/b$ = 3. It implies that the SASR can be extended for the investigation of the τ-periodicity. For base sequence $X = \{x_t \mid t = 1, 2, \ldots, N\}$, the τ-PM of the posterior sub-sequence at position t is generated by the following recurrence formula:

$$M_\Lambda(P_X(t)) = \begin{cases} M_\Lambda(P_X(t+1)) >> b & x_{t+1} \neq \Lambda \\ M_\Lambda(P_X(t+1)) >> b + M^b & x_{t+1} = \Lambda \end{cases} \quad (5\text{-}4)$$

$$M_\Lambda(P_X(N)) = 0$$

Here, M^b denotes a τ-periodicity vector $\{m_1, m_2, \ldots, m_a\}$, in which $m_j = 1$ when $j = b$ and $m_j = 0$ elsewhere.

A τ-periodicity sequence $S(X) = \{s_t \mid t = 1, 2, \ldots, N\}$ can be generated by selecting the row vector as in Equation (3-5). An example is shown in Figure 5.1. And the τ-periodicity walk $W = \{w_t \mid t = 1, 2, \ldots, N\}$ is generated by Equation (3-7) in the τ-periodicity vector space and Equation (3-8) in the complex plane.

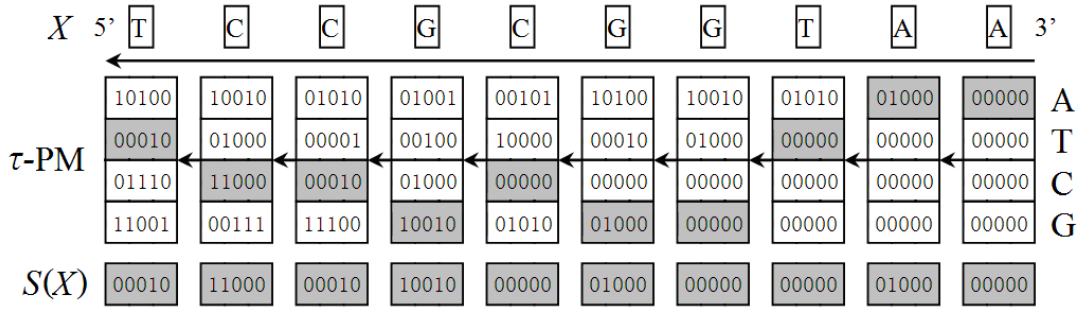


Figure 5.1 A sketch of the algorithm to generate a τ-periodicity sequence. The parameters in this sample: $a = 5$, $b = 2$, $\tau = a/b = 2.5$.

The behavior of a τ-periodicity walk is similar to those of TP walks. The walk of a simple sequence with a high intensity of τ-periodicity should have a trend to move rightward, while the walk of a non τ-periodicity sequence is uniformly random. Similar to the discussion in Section 3.2.1, this principle can be roughly proved as follows.

Consider a DNA sequence X with a τ-PM = $\{M_A(X), M_T(X), M_C(X), M_G(X)\}^T$, where $\tau = a/b$. Also assume that most of the posterior subsequences share the same entries' proportions of the τ-PM only with a shift caused by the position value t. Then, for each step t, the increment in Equation (3-7) is:

$$\frac{s_t}{L(s_t)} = \frac{M_{x_t}(P_X(t))}{L(M_{x_t}(P_X(t)))} \approx \frac{M_{x_t}(X) << bt}{L(M_{x_t}(X))}$$

According to the definition of τ-PM, i.e., Equation (5-2), a certain base Λ appears in the sequence with a probability:

$$\begin{aligned} &\Pr\{x_t = \Lambda \text{ and } bt \,\%\, a = j\} \\ &= \Pr\{x_t = \Lambda \text{ and } b(na + r) \,\%\, a = j\} \\ &= \Pr\{x_t = \Lambda \text{ and } br \,\%\, a = j\} \\ &= \frac{m_{\Lambda j}}{N} \end{aligned}$$

Hence, for each step t, the increment of the TP walk is expected to be:

$$\begin{aligned} E\left(\frac{s_t}{L(s_t)}\right) \approx E\left(\frac{M_{x_t}(X) << bt}{L(M_{x_t}(X))}\right) &= \sum_{\Lambda=A,T,C,G} \sum_{j=1}^{a} \frac{m_{\Lambda,j}}{N} \cdot \frac{M_\Lambda(X) << j}{L(M_\Lambda(X))} \\ &= \sum_{\Lambda=A,T,C,G} \frac{\sum_{j=1}^{a} m_{\Lambda,j} \cdot \{m_{\Lambda,[(j+1)\%a]}, m_{\Lambda,[(j+2)\%a]}, ..., m_{\Lambda,[(j+a-1)\%a]}, m_{\Lambda,j}\}}{N \cdot L(M_\Lambda(X))} \qquad (5\text{-}5) \\ &= \sum_{\Lambda=A,T,C,G} \frac{\{R_\Lambda(1), R_\Lambda(2), ..., R_\Lambda(a)\}}{N \cdot L(M_\Lambda(X))} \end{aligned}$$

$$R_\Lambda(k) = \sum_{j=1}^{a} \left(m_{\Lambda,j} \cdot m_{\Lambda,[(j+k)\%a]}\right) \qquad (k = 1, 2, 3, ..., a)$$

$R_\Lambda(k)$ is the (cyclic) Autocorrelation Function (ACF) of the array $\{m_{\Lambda 1}, m_{\Lambda 2}, \ldots, m_{\Lambda a}\}$. According to the mapping in Equation (5-3), the expected vector in Equation

(5-5) is mapped into the complex plane:

$$E\left(\frac{s_t}{L(s_t)}\right) \to \sum_{\Lambda=A,T,C,G} \frac{\sum_{k=1}^{a}\left(e^{-i\frac{2k\pi}{a}} \cdot R_\Lambda(k)\right)}{N \cdot L(M_\Lambda(X))} = \sum_{\Lambda=A,T,C,G} \frac{p_\Lambda\left(\frac{1}{a}\right)}{N \cdot L(M_\Lambda(X))} \qquad (5\text{-}6)$$

According to Wiener-Khinchin theorem (Ricker, 2003), $p_\Lambda(f)$ is the Power Spectral Density function (PSD) of the array $\{m_{\Lambda 1}, m_{\Lambda 2}, \ldots, m_{\Lambda a}\}$, which is real and non-negative. For a random sequence, the periodicity profile (the array) is close to uniformly constant, i.e., $m_{\Lambda i} \approx m_{\Lambda j}$ for any i and j, so $p_\Lambda(1/a) \approx 0$ in Equation (5-6) and the walk appears random around the zero point. On the other hand, for a sequence with the τ-periodicity, the periodicity profile (the array) should be biased so as to obtain a high norm value of the Fourier spectrum in Equation (5-3). Therefore, the PSD of the array (with a length of a) at the frequency $1/a$ is positive, i.e., $p_\Lambda(1/a) > 0$ in Equation (5-6). It causes the τ-periodicity walk to move rightward in the complex plane.

Similar to the discussions in Section 3.2.2 and Section 3.2.3, it is also found that, the τ-periodicity walk of a complete sequence, consisting of some τ-periodicity and non τ-periodicity regions, has local moving trends. And the corners between neighboring τ-periodicity regions follow the "τ-periodicity corner rule": The corner's angel α (from the previous trend to the posterior trend) can be expressed by $\alpha = -2\Delta\pi/\tau$, with Δ indicating the "frame shift" between the neighboring regions.

5.1.2 The SASR for dinucleotide sequences

As mentioned in Section 2.5, the ~10bp periodicity, related to nucleosome positioning, exists in dinucleotide sequences. Hence, it is significant to develop an SASR approach for dinucleotide sequences to investigate such a periodicity property.

The dinucleotide sequence, corresponding to a given base sequence $X = \{x_t \mid t = 1, 2, \ldots, N\}$, is presented as $Y = \{y_t \mid t = 1, 2, \ldots, N\text{-}1\}$, where y_t is the dinucleotide "x_t - x_{t+1}" (Figure 5.2a). Since there are 4 kinds of different nucleotide bases (Λ = A, T, G, or C), the total number of different dinucleotides Φ is 16 (4^2). To investigate the τ-periodicity property ($\tau = a/b$) in this dinucleotide sequence, a 16×a matrix is built as the τ-periodicity matrix (Figure 5.2b). Each row in the matrix is a τ-periodicity vector M_Φ for dinucleotide Φ, and the τ-PM of the posterior subsequence of the dinucleotide sequence at position t ($1 \le t \le N\text{-}1$) is generated from the following recurrence formula:

$$M_\Phi(P_Y(t)) = \begin{cases} M_\Phi(P_Y(t+1)) >> b & y_{t+1} \neq \Phi \\ M_\Phi(P_Y(t+1)) >> b + M^b & y_{t+1} = \Phi \end{cases} \qquad (5\text{-}7)$$
$$M_\Phi(P_Y(N-1)) = 0$$

A τ-periodicity sequence can also be generated by selecting the row vectors from the τ-PM. However, in some cases, nucleosome positioning as an instance, only some certain kinds of dinucleotides are significant to be considered, while others can be ignored. For this reason, a dinucleotide set L is built containing the concerned kinds of dinucleotides. And the τ-periodicity sequence $S(Y) = \{s_t \mid t = 1, 2, \ldots, N\text{-}1\}$ is generated by

$$s_t = \begin{cases} M_{y_t}(P_Y(t)) & y_t \in L \\ 0 & y_t \notin L \end{cases} \tag{5-8}$$

After that, a τ-periodicity walk can be defined as Equation (3-7) in the τ-periodicity vector space and Equation (3-8) in the complex plane.

t	X	Y	bt % a
1	A	A-T	2
2	T	T-C	4
3	C	C-A	1
4	A	A-T	3
5	T	T-G	5
6	G	G-G	2
7	G	G-T	4
8	T	T-C	1
9	C	C-T	3
10	T	T-G	5
11	G	G-G	2
12	G	G-C	4
13	C	C-A	1
14	A	A-T	3
15	T	T-T	5
16	T		

(a)

	1	2	3	4	5
A-A	0	0	0	0	0
A-T	0	1	2	0	0
A-C	0	0	0	0	0
A-G	0	0	0	0	0
T-A	0	0	0	0	0
T-T	0	0	0	0	1
T-C	1	0	0	1	0
T-G	0	0	0	0	2
C-A	2	0	0	0	0
C-T	0	0	1	0	0
C-C	0	0	0	0	0
C-G	0	0	0	0	0
G-A	0	0	0	0	0
G-T	0	0	0	1	0
G-C	0	0	0	1	0
G-G	0	2	0	0	0

(b)

Figure 5.2 A dinucleotide sequence and the τ-periodicity matrix of the dinucleotide sequence (a) The dinucleotide sequence Y of a nucleotide base sequence X (b) The τ-periodicity matrix of the dinucleotide sequence Y, with $a = 5$, $b = 2$, $\tau = a/b = 2.5$

5.1.3 Investigation of the relationship between sequence periodicity and nucleosome formation

Record files of the nucleosome positioning along chromosome DNA sequences of *Caenorhabditis elegans* are downloaded from UCSC Genome Bioinformatics Site (http://moma.ki.au.dk/genome-mirror/). The files record the number (score) n_j of

nucleosome binding occurrences at each interval j from position α_j to β_j, according to 50bp experimental short reads aligned to the *C. elegans* chromosome sequences (in both DNA strands), as mentioned in Section 2.3.2. Therefore, a spectrum $f(t)$ can be generated to describe the nucleosome binding occurrence at each sequence position t:

$$f(t) = \sum_{j=1}^{O} n_j \cdot \delta(j,t) \tag{5-9}$$

Here, O is the number of recorded intervals and

$$\delta(j,t) = \begin{cases} 1 & \alpha_j \le t \le \beta_j \\ 0 & elsewhere \end{cases} \tag{5-10}$$

Besides, with consideration of the distribution of the nucleosome binding in the τ period ($\tau \approx 10.2$bp for *C. elegans* as previously reported (Widom, 1996; Dlakic et al., 2005)), another spectrum $h(t)$ is presented here, by modifying Equation (5-9) to:

$$h(t) = \left| \sum_{j=1}^{O} n_j \cdot \delta(j,t) \cdot e^{i\frac{2(t-\alpha_j)\pi}{\tau}} \right| \tag{5-11}$$

Figure 5.3a is the smoothed plot of the spectrum $f(t)$ for the fragment from position 100,501 to 104,500 in the 2nd *C. elegans* chromosome sequence (GenBank no. NC_003280) and the smoothed spectrum $h(t)$ for the same fragment is plotted in Figure 5.3b. Compared with $f(t)$, $h(t)$ describes not only the nucleosome binding occurrence, but also the binding preference in the τ period: A high peak in Figure 5.3b needs a high preference for a certain phase in the τ period. Such preference might be related to some sequence-specific periodical signal, such as the 10.2bp periodicity in dinucleotide sequences as reported previously (Widom, 1996; Dlakic et

al., 2005).

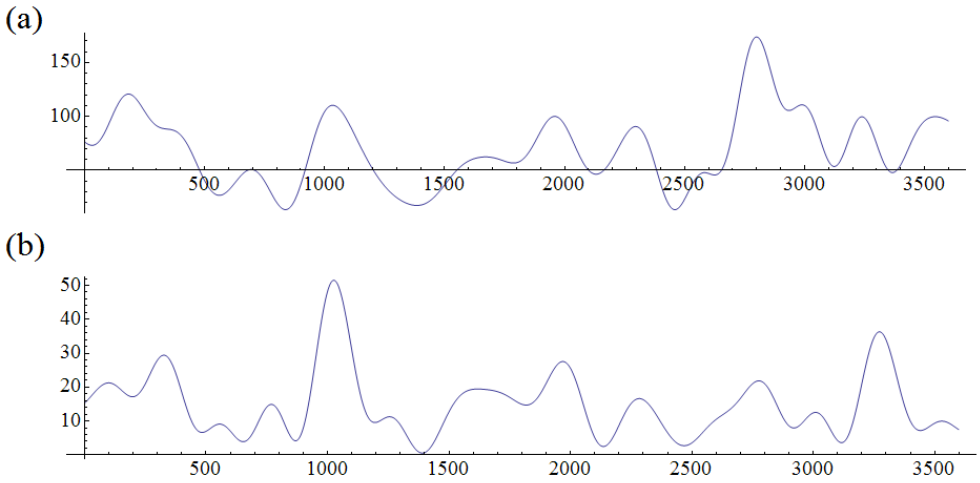


Figure 5.3 Spectrums of (a) the nucleosome binding $f(t)$ and (b) the phase-preferred nucleosome binding $h(t)$

In order to investigate the relationship between the periodicity of dinucleotides and nucleosome binding, the τ-periodicity dinucleotide SASR is adopted. For τ = 10.2, a = 51 and b = 5. Firstly, only one of the 16 dinucleotides is considered each time and set into the dinucleotide set L. Therefore, for each dinucleotide set L containing dinucleotide d, i.e., $L = \{d\}$, a Rightward Rate spectrum $r_L(t)$ can be generated, by calculating the RR values (see Section 3.3.1) of the local sequences in a slide window with a fixed length of N_W = 148bp (the length of a nucleosome for *C. elegans*). Alternatively, the RR value is calculated for a dinucleotide sequence as:

$$RR = \frac{1}{J}\max\{\mathrm{Re}(w_t) \mid t = 1, 2, ..., N_W\}$$

Here, J stands for the number of the concerned dinucleotides in the total N_W dinucleotides. After that, $r_L(t)$ is compared with the spectrum $h(t)$, and the correlation

coefficient $\rho(L, h)$ is calculated by:

$$\rho(L,h) = \frac{N\sum_{t=1}^{N} h(t)r_L(t) - \sum_{t=1}^{N} r_L(t)\sum_{t=1}^{N} h(t)}{\sqrt{N\sum_{t=1}^{N} h^2(t) - \left(\sum_{t=1}^{N} h(t)\right)^2}\sqrt{N\sum_{t=1}^{N} r_L{}^2(t) - \left(\sum_{t=1}^{N} r_L(t)\right)^2}} \tag{5-12}$$

Here, N is the length of the sequence. Table 5.1 lists ρ for all 16 possible dinucleotide sets L.

Table 5.1 The correlation coefficients for the 16 dinucleotide sets L, each containing only 1 dinucleotide

L	{AA}	{AT}	{AG}	{AC}	{TA}	{TT}	{TG}	{TC}
$\rho(L, h)$	0.164	0.334	0.374	-0.332	0.081	0.480	0.472	-0.069
L	{GA}	{GT}	{GG}	{GC}	{CA}	{CT}	{CG}	{CC}
$\rho(L, h)$	-0.147	-0.092	0.036	0.277	0.078	-0.081	0.199	0.078

Furthermore, the dinucleotide sets L, each containing 2 different dinucleotides, are considered. RR spectrums are generated for such sets L (120 combinations overall), and the correlation coefficients are listed in Table 5.2. The spectrum $r_L(t)$ of L = {AG, TG} is found to be the most positively correlated with $h(t)$, according to Table 5.2. Figure 5.4a and Figure 5.4b show that, when previously considering AG and TG separately, some of the nucleosome bindings (with phase preference) are related to the periodicity of AG (in green rectangles), while some are related to that of TG (in red rectangles). The ρ values for AG and TG are less than that for TT when considering each separately, but the combination of these two dinucleotides produces

a higher correlation (Figure 5.4c) than that from any L containing TT. It implies that in order to find out the dinucleotide set L (containing any number of dinucleotides), whose periodicity is mostly related to the "phase-preferred" nucleosome binding, all the 2^{16} combinations need to be tested. For this time-consuming optimization problem, an approximate but fast solution is developed based on the Genetic Algorithm (GA).

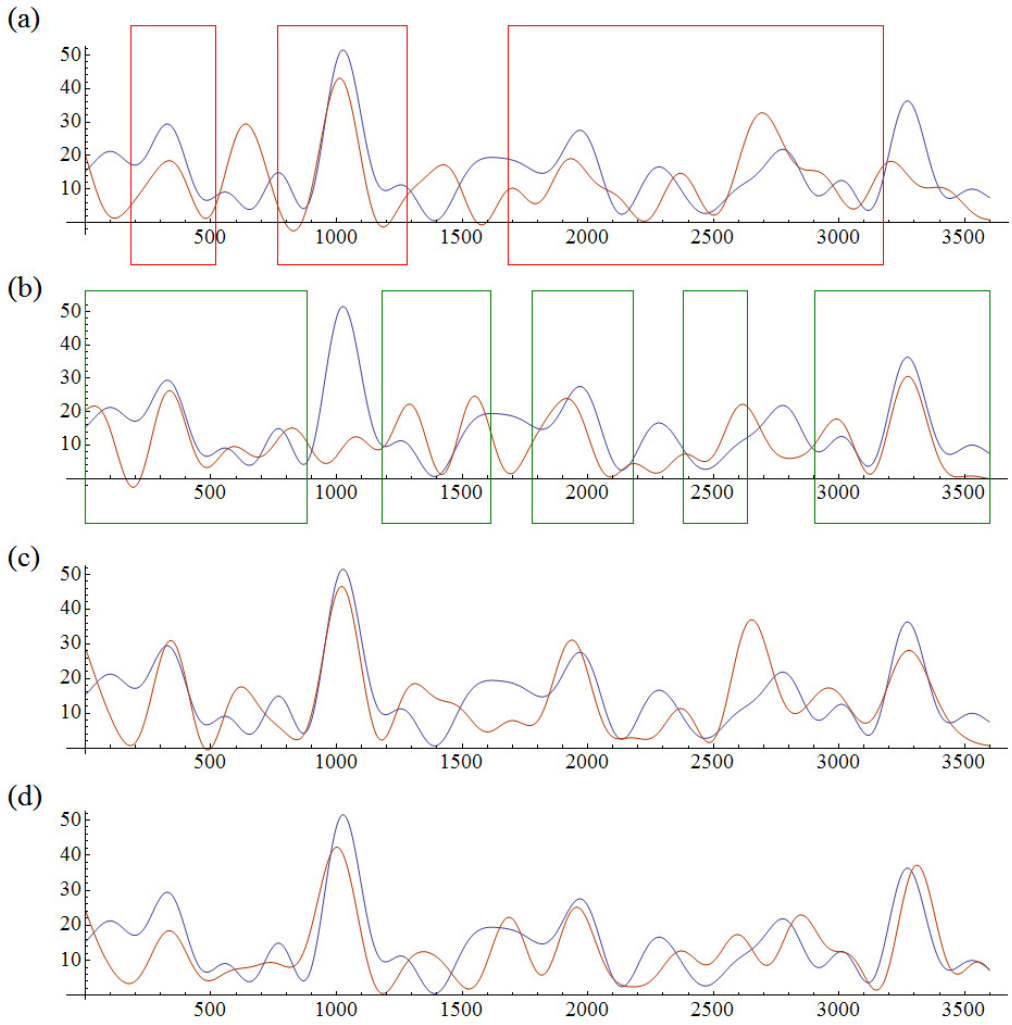


Figure 5.4 Plot of $r_L(t)$ (red) and $h(t)$ (blue) (a) L = {TG} (b) L = {AG} (c) L = {AG, TG} (d) L = {AT, AG, TA, TG, CG}

Table 5.2 The correlation coefficients for the 120 dinucleotide sets L, each containing 2 dinucleotides

L	{AA, AT}	{AA, AG}	{AA, AC}	{AA, TA}	{AA, TT}	{AA, TG}
$\rho(L, h)$	0.418	0.273	-0.001	0.050	0.514	0.481
L	{AA, TC}	{AA, GA}	{AA, GT}	{AA, GG}	{AA, GC}	{AA, CA}
$\rho(L, h)$	0.076	-0.006	-0.037	0.092	0.471	0.264
L	{AA, CT}	{AA, CG}	{AA, CC}	{AT, AG}	{AT, AC}	{AT, TA}
$\rho(L, h)$	0.030	0.377	0.300	0.475	0.115	0.273
L	{AT, TT}	{AT, TG}	{AT, TC}	{AT, GA}	{AT, GT}	{AT, GG}
$\rho(L, h)$	0.486	0.556	0.276	0.260	0.172	0.437
L	{AT, GC}	{AT, CA}	{AT, CT}	{AT, CG}	{AT, CC}	{AG, AC}
$\rho(L, h)$	0.489	0.410	0.180	0.452	0.400	0.124
L	{AG, TA}	{AG, TT}	{AG, TG}	{AG, TC}	{AG, GA}	{AG, GT}
$\rho(L, h)$	0.258	0.604	0.660	0.296	0.169	0.113
L	{AG, GG}	{AG, GC}	{AG, CA}	{AG, CT}	{AG, CG}	{AG, CC}
$\rho(L, h)$	0.252	0.608	0.444	0.294	0.490	0.515
L	{AC, TA}	{AC, TT}	{AC, TG}	{AC, TC}	{AC, GA}	{AC, GT}
$\rho(L, h)$	-0.187	0.380	0.118	-0.230	-0.244	-0.296
L	{AC, GG}	{AC, GC}	{AC, CA}	{AC, CT}	{AC, CG}	{AC, CC}
$\rho(L, h)$	-0.240	-0.013	-0.143	-0.311	-0.182	-0.213
L	{TA, TT}	{TA, TG}	{TA, TC}	{TA, GA}	{TA, GT}	{TA, GG}
$\rho(L, h)$	0.526	0.494	-0.087	-0.079	-0.154	0.113
L	{TA, GC}	{TA, CA}	{TA, CT}	{TA, CG}	{TA, CC}	{TT, TG}
$\rho(L, h)$	0.557	0.226	-0.011	0.248	0.164	0.534
L	{TT, TC}	{TT, GA}	{TT, GT}	{TT, GG}	{TT, GC}	{TT, CA}
$\rho(L, h)$	0.471	0.466	0.468	0.503	0.514	0.428
L	{TT, CT}	{TT, CG}	{TT, CC}	{TG, TC}	{TG, GA}	{TG, GT}
$\rho(L, h)$	0.476	0.515	0.466	0.408	0.398	0.306
L	{TG, GG}	{TG, GC}	{TG, CA}	{TG, CT}	{TG, CG}	{TG, CC}
$\rho(L, h)$	0.487	0.549	0.322	0.304	0.421	0.436
L	{TC, GA}	{TC, GT}	{TC, GG}	{TC, GC}	{TC, CA}	{TC, CT}
$\rho(L, h)$	-0.143	-0.094	0.033	0.349	0.108	-0.110
L	{TC, CG}	{TC, CC}	{GA, GT}	{GA, GG}	{GA, GC}	{GA, CA}
$\rho(L, h)$	0.097	0.058	-0.174	-0.090	0.306	0.067
L	{GA, CT}	{GA, CG}	{GA, CC}	{GT, GG}	{GT, GC}	{GT, CA}
$\rho(L, h)$	-0.247	0.020	0.077	-0.042	0.333	0.170
L	{GT, CT}	{GT, CG}	{GT, CC}	{GG, GC}	{GG, CA}	{GG, CT}
$\rho(L, h)$	-0.047	0.173	0.149	0.315	0.157	0.004
L	{GG, CG}	{GG, CC}	{GC, CA}	{GC, CT}	{GC, CG}	{GC, CC}
$\rho(L, h)$	0.325	0.077	0.280	0.381	0.378	0.211
L	{CA, CT}	{CA, CG}	{CA, CC}	{CT, CG}	{CT, CC}	{CG, CC}
$\rho(L, h)$	0.041	0.167	0.095	0.188	-0.015	0.201

The designed GA-based method represents every possible dinucleotide set L by a chromosome. It is aimed at maximizing the objective function, i.e., the correlation coefficient $\rho(L, h)$, by finding out the optimal dinucleotide set L.

The flow of the method is presented as follows (Figure 5.5). Initially, a population of chromosomes is set up randomly. And then the evolution takes place in this population iteratively for several generations. In each generation of the evolution, different from commonly used GA-based methods, in this study, all the chromosomes in the population are selected, and a crossover step is processed on all these chromosomes, followed by a mutation step. After both the crossover and the mutation steps, the chromosomes in the population are sorted according to the objective function, i.e., $\rho(L, h)$. Chromosomes with low values of $\rho(L, h)$ are eliminated and the size of the population remains the same as that of the previous generation. Then the evolution moves on to the next generation. It must be noticed that, in this cycle of evolution, if the chromosomes in the population become the same and this convergent state lasts for 7 generations, a "regeneration" process should be conducted by keeping only one chromosome and replacing the rest with random chromosomes. Finally, if the population keeps in a convergent state, and the best chromosome with the highest value of $\rho(L, h)$ has kept unchanged for a long period (> 7 generations), this chromosome is outputted as an optimized result.

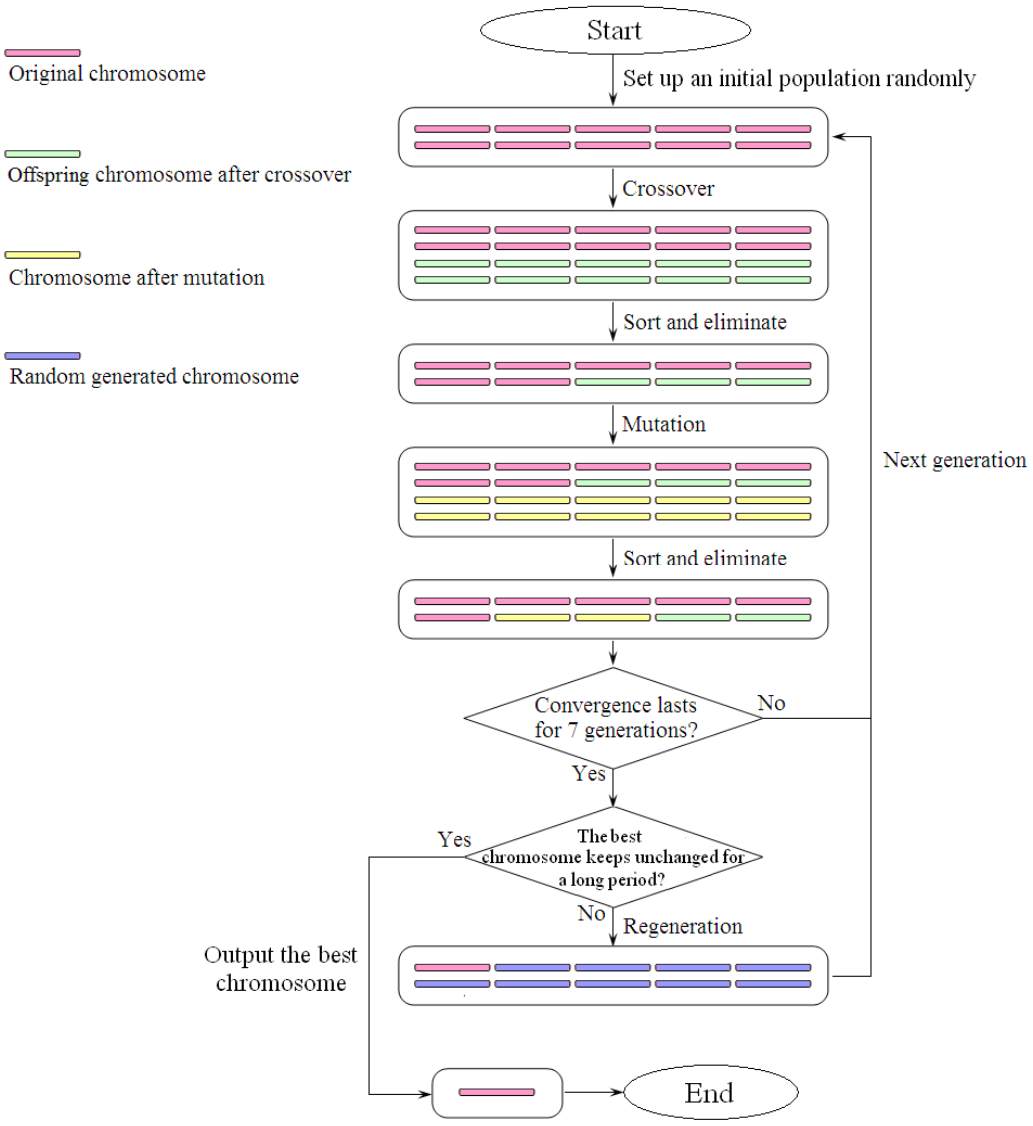


Figure 5.5 The flow of the GA-based method

As shown in Figure 5.6a, in this designed GA-based method, the dinucleotide set L is represented as a binary array (a chromosome in this GA-based method), i.e., $C = \{c_0, c_1, \ldots, c_{15}\}$, in which the 16 binaries correspond to the 16 dinucleotides AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, and CC, respectively. $c_d = 1$ ($d = 0, 1, 2, \ldots, 15$) indicates that the corresponding dinucleotide d is

contained in the dinucleotide set L.

In the designed crossover step (Figure 5.6b), all the chromosomes are paired randomly and the (single point) crossover of each pair generates two offspring by swapping the sub-strings after a random site between these two parents. In the mutation step (Figure 5.6c), each chromosome has an opportunity to mutate on at maximum 3 random sites. As shown in Figure 5.6c, the mutation on a certain site is to change the value into its opposite ($0 \leftrightarrow 1$). After that, the original chromosome is kept in the population, instead of being replaced as in commonly used GA-based methods.

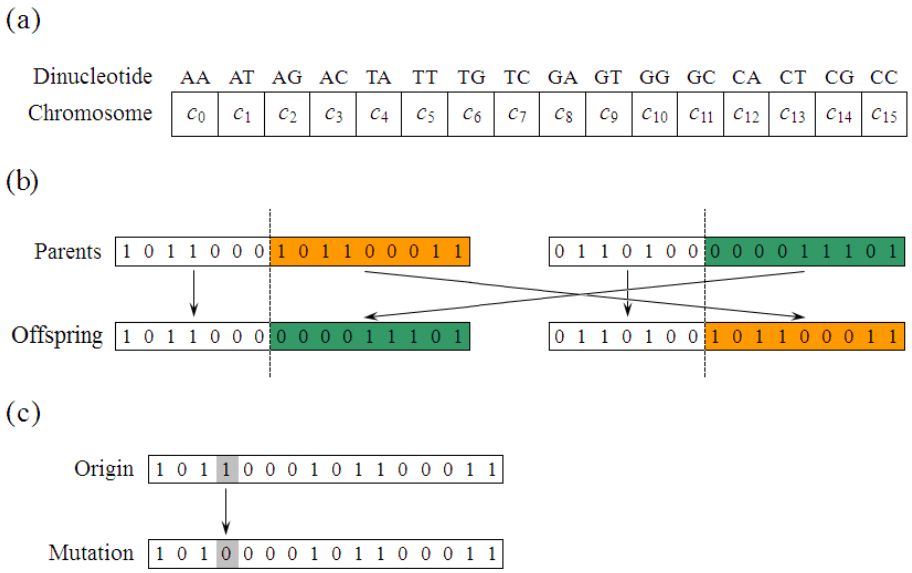


Figure 5.6 The GA-based method (a) The designed chromosome (b) The (single point) crossover (c) The mutation

For the sample DNA fragment mentioned previously, an initial population with

a size of 10 is randomly generated and the designed GA-based method is applied to this population. After 30 generations, a convergent optimization is obtained as C = {0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0}, representing the dinucleotide set L = {AT, AG, TA, TG, CG}. Figure 5.4d is the smoothed plot of $r_L(t)$, compared with $h(t)$. It shows that in most places of the sequence, the trend of $r_L(t)$ fits well to $h(t)$. The correlation coefficient of these two curves is 0.702, suggesting that the 10.2bp periodicity of the dinucleotides {AT, AG, TA, TG, CG} is highly related to the phase-preferred nucleosome binding in this sample DNA fragment.

To investigate the significance of the obtained value of ρ, 100 random sequences with a length of 4,000bp (the same as that of the sample DNA fragment given previously) are generated. The designed GA algorithm is applied to each of these 100 sequences to find the dinucleotide combination L producing $r_L(t)$ mostly related to $h(t)$, and the maximal value of $\rho(L, h)$ is recorded as $\rho_{\max}$. In the distribution of $\rho_{\max}$ achieved by these random sequences (Figure 5.7), only 23 out of 100 sequences (23%) reach 0.5, 9 sequences get $\rho_{\max} > 0.6$, and none of them could achieve the value of 0.7. It indicates that the correlation coefficient (0.702) from the real DNA fragment described previously has high statistic significance and is not obtained by chance, implying that the 10.2bp periodicity of the dinucleotides {AT, AG, TA, TG, CG} in the given real DNA fragment do contain information related to the "phase-preferred" nucleosome binding.

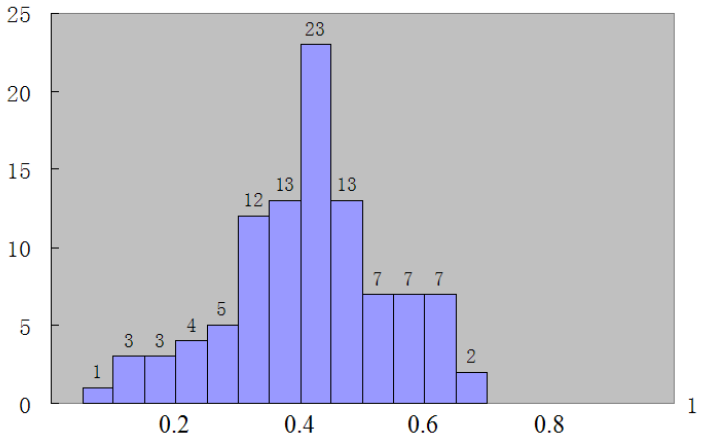


Figure 5.7 Plot of the distribution of ρ_{max} achieved by random sequences

Furthermore, the nucleosome data are collected for 9 more fragments (10 fragments together with the previous one), which are randomly selected from non-featured regions in *C. elegans* chromosomes (GenBank no. NC_003279 ~ NC_003283). For each fragment, a spectrum $h(t)$ is generated, and the dinucleotide set L, whose 10.2bp periodicity is mostly related to the phase-preferred nucleosome binding, is found by the GA algorithm mentioned above. The results are listed in Table 5.3. It clearly shows that, when comparing the nucleosome data with the corresponding real DNA sequences, the distribution of ρ_{max} is obviously different from that in the random cases (Figure 5.7). All the 10 fragments achieve $\rho_{max} > 0.5$ and there are even 2 fragments with ρ_{max} beyond 0.7. This result confirms that the relationship between the phase-preferred nucleosome binding and the 10.2bp periodicity of certain dinucleotide combination exists in chromosome DNA comprehensively.

Table 5.3 Maximum value of $\rho(L, h)$ for sample fragments from the *C. elegans* chromosomes and the optimized dinucleotide sets L obtained by the GA-based method

Fragment	C_{opt}																ρ_{max}
	AA	AT	AG	AC	TA	TT	TG	TC	GA	GT	GG	GC	CA	CT	CG	CC	
ChrI 122,001 ~ 125,000	1	0	0	0	1	0	0	1	0	0	0	1	0	1	1	0	0.613
ChrI 2,151,001 ~ 2,154,000	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0.655
ChrII 100,501 ~ 104,500	0	1	1	0	1	0	1	0	0	0	0	0	0	0	1	0	0.702
ChrII 3,954,001 ~ 3,957,000	1	0	0	1	0	0	0	1	0	1	1	1	1	1	0	1	0.591
ChrIII 9,879,001 ~ 9,882,000	0	0	0	0	1	0	0	1	0	1	0	0	0	0	1	1	0.520
ChrIII 2,868,001 ~ 2,871,000	0	0	1	0	1	1	0	0	1	0	0	1	0	1	0	0	0.596
ChrIV 1,155,001 ~ 1,158,000	0	0	0	1	1	1	0	0	0	0	1	0	0	0	0	1	0.711
ChrIV 1,281,001 ~ 1,284,000	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0.654
ChrV 6,240,001 ~ 6,243,000	0	0	0	1	0	0	0	0	0	1	0	1	0	1	1	0	0.552
ChrV 1,200,001 ~ 1,203,000	0	1	0	1	1	0	0	0	0	1	1	1	0	0	1	0	0.543
Score	1.204	1.245	1.298	2.397	4.340	1.307	0.702	1.724	0.596	3.515	1.845	3.549	1.246	2.352	3.584	1.822	

To investigate the usage of each dinucleotide in the optimized dinucleotide sets, a score for each dinucleotide is calculated: each fragment in this experiment "votes" for the dinucleotides contained in its optimized dinucleotide set with the voting weight of ρ_{max} and the summation of the votes is the score for each dinucleotide (Table 5.3). As a result, dinucleotide "TA" gets the highest score of 4.340. It implies that the 10.2bp periodicity of "TA" is related to the phase-preferred nucleosome binding in the most extensive places of the chromosomes. This result conforms to the finding from Takasuka and Stein (2010). The 10.2bp periodicity of dinucleotides "GT", "GC", and "CG" are also found to be related to the binding very comprehensively, with scores of 3.515, 3.549, and 3.584, respectively. The high

score of “GC” conforms to previous reports (Segal et al., 2006; Chen et al., 2010), while the relationship between the binding and the periodicity of “GT” and “CG” has seldom been reported in previous literatures. On the other hand, dinucleotide “GA” has the lowest score of 0.596. It reveals that its periodicity has no significant relationship with the phase-preferred nucleosome binding, and actually, it was also not reported in any previous literature. It must be noticed that the dinucleotides with medium scores are not excluded from being related to nucleosome binding. Like in Figure 5.4, the medium-scored dinucleotides might connect their 10.2bp periodicity with nucleosome formation only in some local regions. Instead of aligning all nucleosomes to center or beginning (Segal et al., 2006; Chen et al., 2008; Chen et al., 2010), the comparison between a nucleosome formation profile, e.g. $h(t)$, and an RR spectrum along a DNA sequence allows for the revelation of such a local relationship between dinucleotide periodicity and nucleosome formation.

The GA method is also applied to investigate the relationship between $r_L(t)$ and $f(t)$. As shown in Table 5.4, the optimal dinucleotide sets L are also found to achieve high correlation coefficients $\rho(L, f)$. However, for nearly all the sample fragments (except for the 5th fragment), the coefficients are lower than the values for $h(t)$. This implies that the 10.2bp periodicity in dinucleotide sequences is related to not only the occurrence of nucleosome formation, but also the binding preference for the phase in the 10.2bp period.

Table 5.4 Maximum values of $\rho(L, f)$ for sample fragments from the *C. elegans* chromosomes, compared with the maximum values of $\rho(L, h)$ for the same fragments

	Fragment	ρ_{max} in all $\rho(L, f)$	ρ_{max} in all $\rho(L, h)$
1	ChrI 122,001 ~ 125,000	0.467	0.613
2	ChrI 2,151,001 ~ 2,154,000	0.534	0.655
3	ChrII 100,501 ~ 104,500	0.531	0.702
4	ChrII 3,954,001 ~ 3,957,000	0.542	0.591
5	ChrIII 9,879,001 ~ 9,882,000	0.614	0.520
6	ChrIII 2,868,001 ~ 2,871,000	0.595	0.596
7	ChrIV 1,155,001 ~ 1,158,000	0.577	0.711
8	ChrIV 1,281,001 ~ 1,284,000	0.640	0.654
9	ChrV 6,240,001 ~ 6,243,000	0.337	0.552
10	ChrV 1,200,001 ~ 1,203,000	0.480	0.543

5.2 Mature SASR and a hypothetical anti-TP property

In this section, the behaviors of TP walks of random sequences are further discussed. An improvement of the original SASR method is then introduced. This improved SASR, i.e., the mature SASR, shows its ability in detecting a hypothetical anti-TP property in DNA sequences.

5.2.1 TP walks of random sequences and a mature SASR

2,000 random sequences are generated with lengths ranging from 300bp to 5,000bp and the SASR is applied to these sequences. The distribution of the SRR values of the TP walks is plotted in Figure 5.8a in the form of its Probability Density

Function (PDF). It shows that the distribution is close to the normal distribution with a slightly negative expected value.

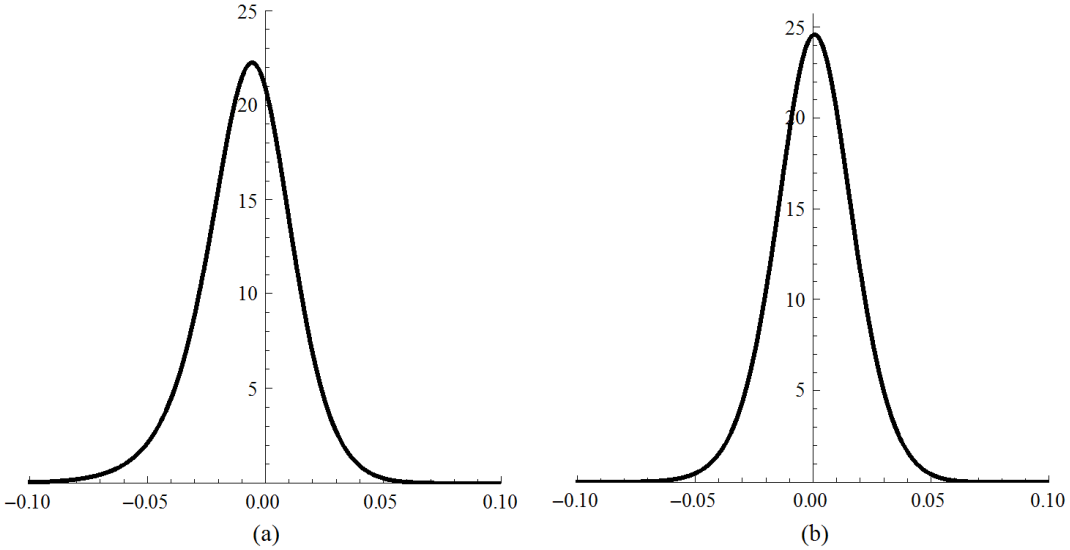


Figure 5.8 The PDF of the distribution of the SRR values when (a) applying the original SASR to the random sequences (b) applying the mature SASR to the random sequences

The reason for the slightly negative distribution is discussed below. Consider a random sequence $X = \{x_t \mid t = 1, 2, 3, \ldots, N\}$. At any position, a certain base Λ (Λ = A, T, C, or G) appears with a fixed probability p_Λ and $p_A + p_T + p_C + p_G = 1$. Suppose a base Λ appears at position t_0, we have:

$$s_{t_0} = M_\Lambda(P_X(t_0)) = \{m_{\Lambda 1}, m_{\Lambda 2}, m_{\Lambda 3}\}$$

$$\text{where} \quad m_{\Lambda j} = \text{count}\{t \mid x_t = \Lambda \ \text{ and } \ (t - t_0)\%\ 3 = j \ \text{ and } \ t > t_0\}$$

It is easy to find that the random variable $m_{\Lambda j}$ follows the Binomial distribution. Use n_j to denote the count of the positions t that satisfy $t > t_0$ and $(t - t_0)\ \%\ 3 = j$. Then:

$$m_{\Lambda j} \sim B(n_j, p_\Lambda)$$

$$\text{The PDF: } f(m_{\Lambda j}) = \binom{n_j}{m_{\Lambda j}} p_\Lambda{}^{m_{\Lambda j}} (1-p_\Lambda)^{n_j - m_{\Lambda j}}$$

$$\text{The expected value: } E(m_{\Lambda j}) = n_j p_\Lambda$$

And the expected value of the TP vector is:

$$E(s_{t_0}) = p_\Lambda \cdot \{n_1, n_2, n_3\}$$

According to the definition of n_j, although the differences among n_1, n_2, and n_3 are no more than 1, n_3 is always the minimum in the three. It causes the walk to move leftward slightly for each step and further produces a slightly negative SRR value.

In order to obtain a normal distribution with an expected value of 0, an alteration of the SASR is proposed, called the mature SASR. In the original SASR, at each position t, the TPM of the posterior subsequence is calculated and the TP vector s_t selected directly from this TPM, as previously mentioned. In this alteration, s_t is selected from a "mature" TPM, instead of from the original matrix. Here, "mature" means that the TPM satisfies:

$$\sum_\Lambda m_{\Lambda 1} = \sum_\Lambda m_{\Lambda 2} = \sum_\Lambda m_{\Lambda 3}$$

A mature TPM is maintained with a simple recurrence formula only involving a RCS: $M_\Lambda(P_X(t)) = M_\Lambda(P_X(t+1)) >> 1$. Besides, the original TPM is still maintained as mentioned before, so that the mature TPM can be updated by copying it, when the original TPM becomes "mature", in every three steps. Figure 5.9 shows a simple example of generating a TP sequence with this new algorithm.

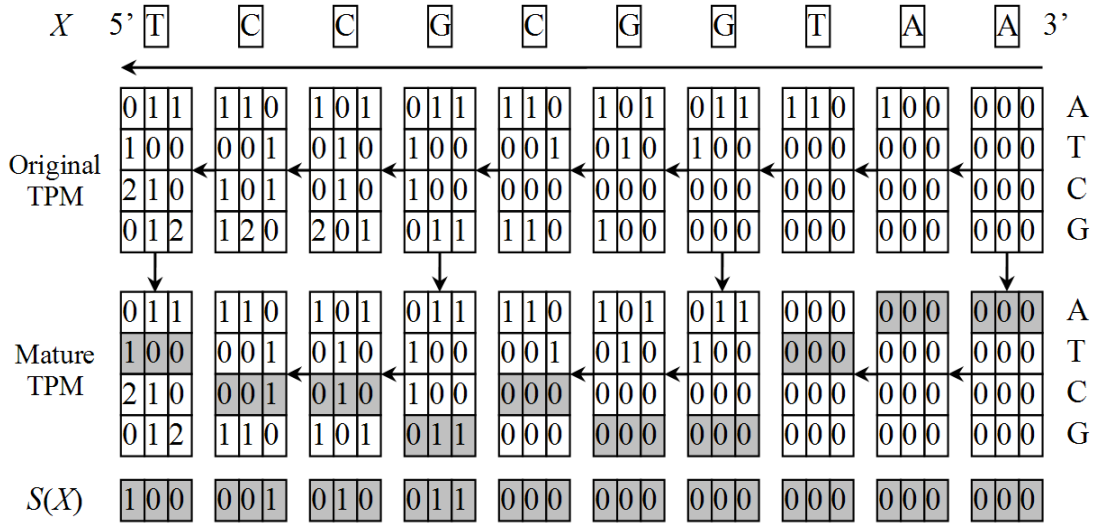


Figure 5.9 A simple example of generating a TP sequence with the new algorithm

This altered SASR is applied to the random sequence dataset and the distribution of the SRR values is plotted in Figure 5.8b. It shows that it is close to the normal distribution with an expected value of 0. So TP walks of random sequences are unbiasedly random around the zero point.

5.2.2 Revealing a hypothetical anti-TP property by the mature SASR

The TP profile was presented in Frenkel and Korotkov's work (Frenkel and Korotkov, 2008; Frenkel and Korotkov, 2009) using a Triplet Periodicity Matrix (TPM) as mentioned in Section 3.1.1. The TP profiles in the parts of a non-TP sequence have no correlation with each other. It shows a "Brownian pattern" in the sequence. On the other hand, in a simple TP sequence $X = \{x_t \mid t = 1, 2, 3, \ldots, N\}$, a certain base Λ appears at position j in the 3bp period with a probability:

$$\Pr\{x_t = \Lambda \text{ and } t\%3 = j\} = \frac{m_{\Lambda j}}{N}$$

Kotlar and Lavner's finding (Kotlar and Lavner, 2003) suggests that, in coding regions of a given organism, the TP profile, by which nucleotide bases appear in the triplet period, tends to keep unchanged. It can be considered as a "persistent pattern" in the sequence:

$$\Pr\{t\%3 = j_0 \mid t \le t' \text{ and } x_t = \Lambda\} = \Pr\{t\%3 = j_0 \mid t > t' \text{ and } x_t = \Lambda\} \quad (5\text{-}13)$$

Besides the "Brownian pattern" and the "persistent pattern" mentioned above, a hypothetical pattern is considered, namely the "anti-persistent pattern", which possibly exists. For the anti-persistent pattern, any part of the sequence has the TP profile opposite to the rest parts. In other words, a certain base Λ avoids appearing at the position *j* in the 3bp period, which is preferred in other parts of the sequence. Contrary to the TP property, here the sequence with such an anti-persistent pattern is said to be with an anti-TP property. To simulate sequences with the anti-TP, an ideal probability model is built here as a simple case. That is: at any given position t_0 % 3 = j_0 in the sequence, a certain base Λ appears with a probability:

$$\Pr\{x_{t_0} = \Lambda\} = \Pr\{x_t = \Lambda \mid t > t_0 \text{ and } t\%3 \neq t_0\%3\}$$

or equivalently, (5-14)

$$\Pr\{x_{t'} = \Lambda \mid t'\%3 = j_0\} = \Pr\{x_t = \Lambda \mid t > t' \text{ and } t\%3 \neq j_0\}$$

So that, for any position *t*' presenting Λ in the sequence,

$$\Pr\{t'\%3 = j_0 \mid x_{t'} = \Lambda\} = \frac{\Pr\{t'\%3 = j_0\} \cdot \Pr\{x_{t'} = \Lambda \mid t'\%3 = j_0\}}{\sum_j \Pr\{t'\%3 = j\} \cdot \Pr\{x_{t'} = \Lambda \mid t'\%3 = j\}}$$

$$\approx \frac{\Pr\{x_t = \Lambda \mid t > t' \text{ and } t\%3 \neq j_0\}}{\sum_j \Pr\{x_t = \Lambda \mid t > t' \text{ and } t\%3 \neq j\}}$$

$$\approx \frac{\Pr\{t\%3 \neq j_0 \mid t > t'\} \cdot \Pr\{x_t = \Lambda \mid t > t' \text{ and } t\%3 \neq j_0\}}{\sum_j \Pr\{t\%3 \neq j \mid t > t'\} \cdot \Pr\{x_t = \Lambda \mid t > t' \text{ and } t\%3 \neq j\}}$$

$$= \Pr\{t\%3 \neq j_0 \mid t > t' \text{ and } x_t = \Lambda\}$$

Therefore, this model is found to be opposite to the "persistent pattern" of Equation (5-13). The discussion on the TP walk's behavior in Section 3.2.1 stands only in cases of the "Brownian pattern" and the "persistent pattern". A sequence with the anti-TP property has a different TP walk trend from coding sequences (moving rightward) or random sequences (moving randomly).

Consider any short section containing three sequential positions t_0 - 2, t_0 – 1, and t_0 (t_0 is a multiple of 3, i.e., t_0 mod 3 = 0) in a sequence with the above probability model. The posterior subsequences at these three positions share a similar TPM with a shift:

$$M_\Lambda(P_X(t_0 - i)) \approx M_\Lambda(P_X(t_0)) >> i$$

Meanwhile, according to Equation (5-14), base Λ appears at these positions with a probability:

$$\Pr\{x_{t_0-i} = \Lambda\} \approx \frac{\sum_{j \neq 3-i} m_{\Lambda j}}{\sum_{j \neq 3-i} m_{\mathrm{A}j} + m_{\mathrm{T}j} + m_{\mathrm{C}j} + m_{\mathrm{G}j}} \qquad (i = 0, 1, 2)$$

Here, $m_{\Lambda j}$ stands for the entry in the TPM of the posterior subsequence at

position t_0. Meanwhile, we have:

$$\sum_{\Lambda} m_{\Lambda 1} \approx \sum_{\Lambda} m_{\Lambda 2} \approx \sum_{\Lambda} m_{\Lambda 3} \approx \frac{N - t_0}{3}$$

Hence, these three steps in the walk move to:

$$\begin{aligned} E\left(\sum_i \frac{s_{t_0-i}}{L(s_{t_0-i})}\right) &\approx E\left(\sum_i \frac{M_{x_{t_0-i}}(P_X(t_0)) >> i}{L(M_{x_{t_0-i}}(P_X(t_0)))}\right) \\ &= \sum_i \sum_{\Lambda=A,T,C,G} \left(\frac{\sum_{j\neq 3-i} m_{\Lambda j}}{\sum_{j\neq 3-i} m_{\mathrm{A}j} + m_{\mathrm{T}j} + m_{\mathrm{C}j} + m_{\mathrm{G}j}} \cdot \frac{M_{\Lambda}(P_X(t_0)) >> i}{L(M_{\Lambda}(P_X(t_0)))} \right) \\ &\approx \frac{3}{2(N-t_0)} \sum_{\Lambda=A,T,C,G} \frac{\sum_i \left[\left(\sum_{j\neq 3-i} m_{\Lambda j} \right) \cdot M_{\Lambda}(P_X(t_0)) >> i \right]}{L(M_{\Lambda}(P_X(t_0)))} \\ &= \frac{3}{2(N-t_0)} \sum_{\Lambda=A,T,C,G} \frac{\{\alpha_1, \alpha_2, \alpha_3\}}{L(M_{\Lambda}(P_X(t_0)))} \end{aligned} \tag{5-15}$$

where,

$$\alpha_1 = \alpha_2 = m_{\Lambda 1} m_{\Lambda 2} + m_{\Lambda 2} m_{\Lambda 3} + m_{\Lambda 3} m_{\Lambda 1} + {m_{\Lambda 1}}^2 + {m_{\Lambda 2}}^2 + {m_{\Lambda 3}}^2$$

$$\alpha_3 = 2 m_{\Lambda 1} m_{\Lambda 2} + 2 m_{\Lambda 2} m_{\Lambda 3} + 2 m_{\Lambda 3} m_{\Lambda 1}$$

Obviously, in this case, we have: $\alpha_1 = \alpha_2 \geq \alpha_3$. Therefore, in Equation (5-15), the first two elements of the expected vector dominate the third one. According to Equation (3-3), it causes the TP walk to move leftward in the complex plane.

In order to verify the capability of the mature SASR in revealing the anti-TP property, the method is applied to a simulated anti-TP dataset. To generate a simulated anti-TP DNA sequence with a length of N, the flow chart in Figure 5.10 is followed. Firstly, a short subsequence at the end (the "seed"), i.e., $\{x_{N-L+1}, x_{N-L+2}, \ldots,$

$x_N\}$, is randomly generated. The TPM of the complete sequence is calculated as follows:

$$m_{\Lambda j} = \text{count}\{t \mid x_t = \Lambda \text{ and } t\%3 = j \text{ and } 1 \le t \le N\}$$

Then the bases in the anterior part are assigned recursively from position N-L to 1. For each given position t, $1 \le t \le N$-L, x_t is assigned to be base Λ with a probability:

$$\Pr\{x_t = \Lambda\} = \frac{\sum_{j \ne t\%3} m_{\Lambda j}}{\sum_{j \ne t\%3} m_{\Lambda j} + m_{\mathrm{T}j} + m_{\mathrm{C}j} + m_{\mathrm{G}j}}$$

After assigning the base at each position, the TPM of the complete sequence is immediately updated, with the newly assigned x_t.

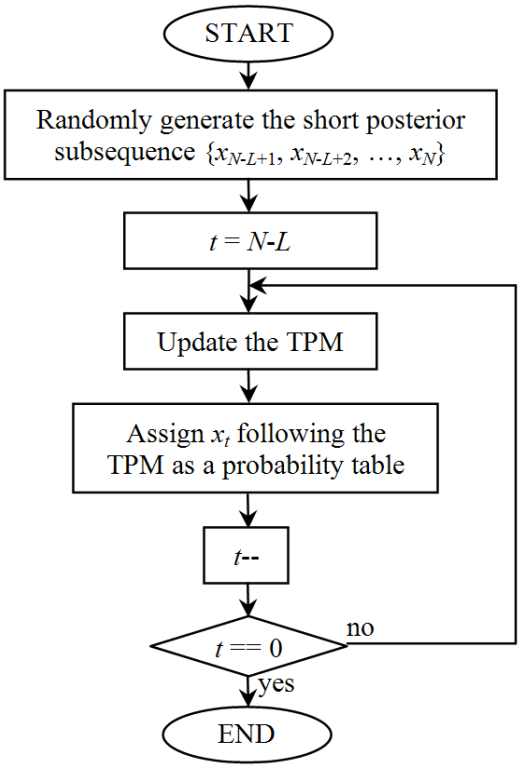


Figure 5.10 The flow chart to generate a simulated anti-TP sequence

2,000 simulated anti-TP sequences with lengths of 300bp ~ 5,000bp are generated with a seed length of $L = 9$. Then the mature SASR is applied to these sequences. The distribution of the SRR values of the TP walks is plotted in Figure 5.11, compared with that for the random sequences. It shows an obvious difference between these two distributions. The PDF curve for the simulated anti-TP sequences is on the left side to that for the random sequences, and the Cumulative Distribution Function (CDF) curves indicate that there are 85% simulated sequences with negative SRR values, while the SRR values of the random sequences distribute fifty-fifty in negative and positive areas. Assuming that the distribution for the random dataset is the normal distribution with an expected value of 0, a *t*-test is conducted to investigate the statistical significance of the anti-TP dataset. It is found that the sample mean and the sample deviation of the 2,000 SRR values for the simulated anti-TP dataset are -0.0157 and 0.0173 respectively. Hence the *t* value is -40.65 and the corresponding *p*-value is near 0, much less than the commonly used threshold, i.e., 5%. This result shows a very high statistical significance of the simulated dataset and reveals that the mature SASR is able to discriminate anti-TP sequences from random sequences. The anti-TP property can be identified according to a leftward moving trend in the TP walk.

Meanwhile, some other methods are also applied to the dataset to check whether they could probably reveal such a hypothetical anti-TP property. 500

sequences from the simulated anti-TP dataset are used to train a Spectral Rotation Measure (SRM) (Kotlar and Lavner, 2003) and then the SRM is applied to the rest 1,500 sequences as well as the random sequences. The distributions of the SRM for the two datasets are plotted in Figure 5.12. There is also a difference between these two distributions. Although they are both close to 0, the values for the anti-TP dataset have a more narrow distribution. It is actually difficult to build a proper test of hypothesis to discriminate the anti-TP sequences from the random ones, since in both cases the SRM tends to 0. However, if a classification is conducted, in which an unknown sequence is classified as anti-TP when its SRM is less than a certain threshold θ, the average values of sensitivity (Sn) and specificity (Sp) can be calculated from the CDF in Figure 5.12b and plotted against the value of θ as in Figure 5.13a. It shows that the highest accuracy is around 62%. On the other hand, the average values of Sn and Sp from such a classification using the SRR can go beyond 69% (Figure 5.13b). Therefore, the mature SASR is said to be with stronger capability of indentifying the hypothetical anti-TP property.

The anti-TP property includes all possible anti-persistent patterns in nucleotide triplets, not limited to the probability model built in this section. In practice, the anti-TP property may be presented in more complicated patterns. Therefore, the mature SASR is also applied to some fragments from real DNA sequences to discover the practical anti-TP patterns. Two examples shown in Figure 5.14 are for the fragment

15,450 ~ 16,565 in the *Odobenus rosmarus* mtDNA (GenBank no. NC_004029) and the fragment 710,139 ~ 710,992 in the 12th chromosome DNA of *S. cerevisiae* (GenBank no. NC_001144) respectively. The TP walks clearly reveal the anti-TP property contained in these two fragments by obvious trends to move leftward. And it is noticed that, the former (from the *Odobenus rosmarus* mtDNA) is annotated as a D-Loop (Kasamatsu et al., 1971), but the latter (from the 12th chromosome DNA of *S. cerevisiae*) is not annotated with any feature. It implies that the anti-TP patterns may carry some information, but the significance and the biological interpretation are unknown so far. Besides dealing with the TP and the ~10bp periodicity, supporting the visualization of the anti-TP property is deemed as another function of the (mature) SASR.

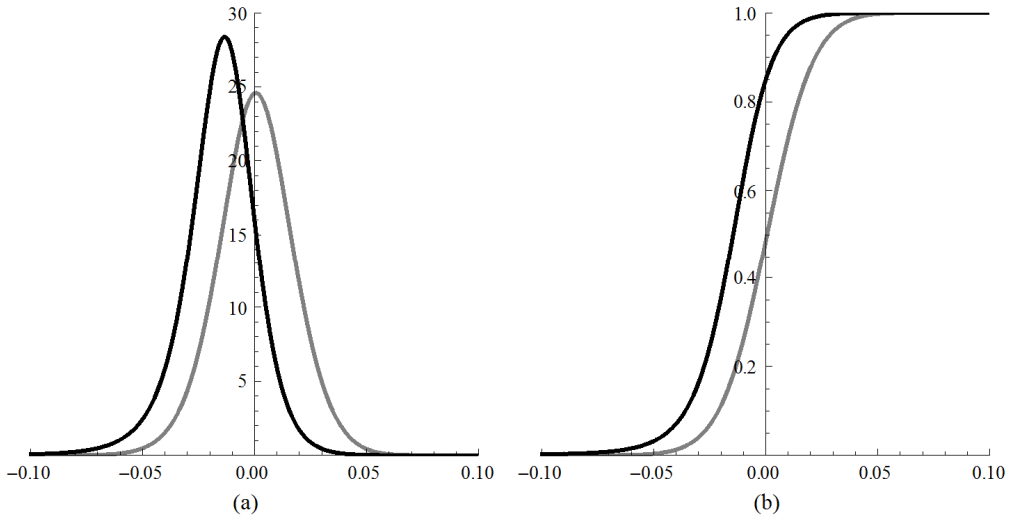


Figure 5.11 The distribution of the SRR values when the mature SASR is applied to the simulated anti-TP sequences (black) compared with that for the random sequences (gray) (a) The Probability Density Function (PDF) (b) The Cumulative Distribution Function (CDF)

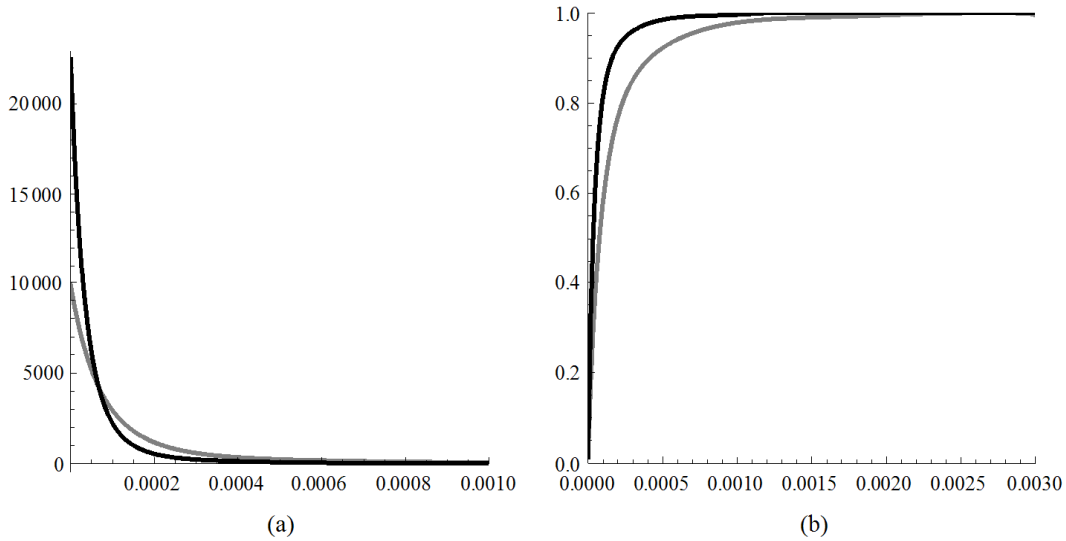


Figure 5.12 The distribution of the SRM applied to the simulated anti-TP sequences (black) compared with that for the random sequences (gray) (a) The Probability Density Function (PDF) (b) The Cumulative Distribution Function (CDF)

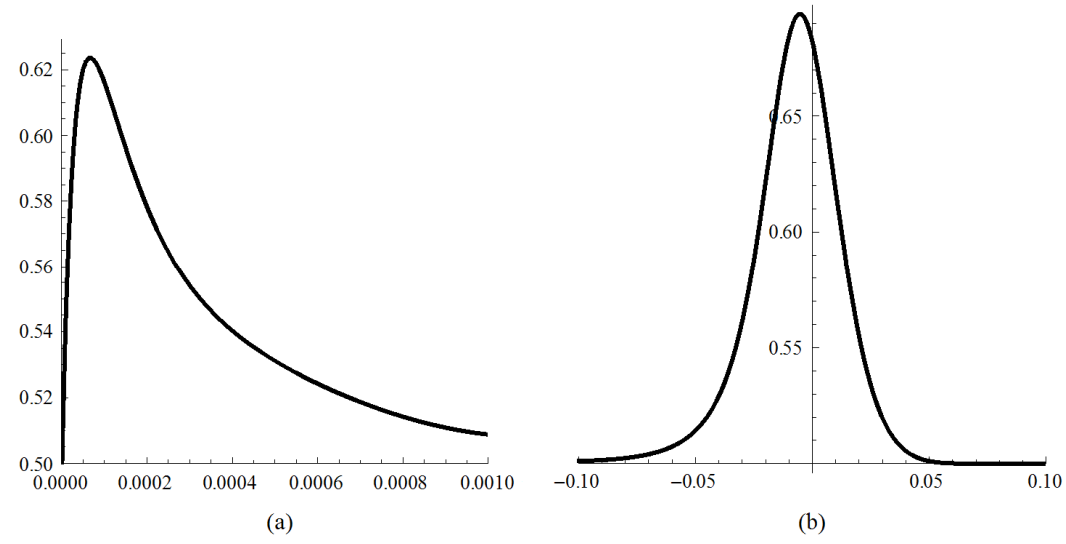


Figure 5.13 The average values of sensitivity (Sn) and specificity (Sp) in determining whether an unknown sequence is anti-TP or random (a) by checking whether the SRM is less than a certain threshold θ (the horizontal axis) (b) by checking whether the SRR is less than a certain threshold θ (the horizontal axis)

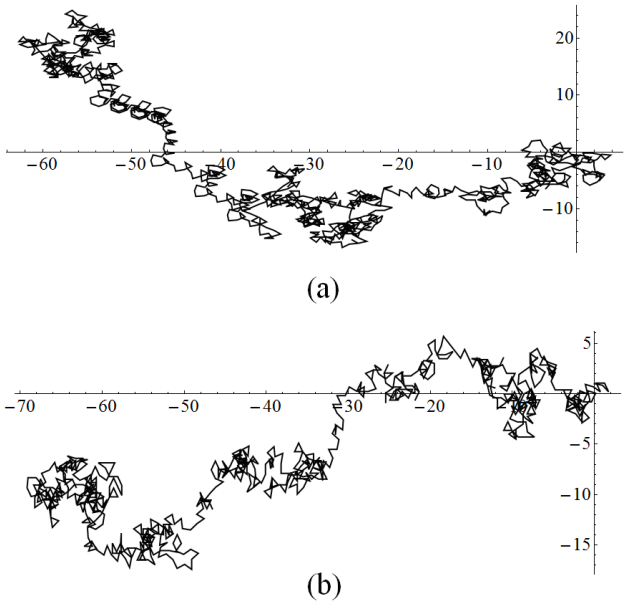


Figure 5.14 Two examples of real DNA fragments containing the anti-TP property, with the TP walks to move leftward (a) Region 15,450 ~ 16,565 in the *Odobenus rosmarus* mtDNA (GenBank no. NC_004029) (b) Region 710,139 ~ 710,992 in the 12th chromosome DNA of *S. cerevisiae* (GenBank no. NC_001144)

5.3 Summary

Some significant extensions of the original SASR have been provided in this chapter. An SASR approach for the τ-periodicity (τ is a rational number) and for dinucleotide sequences has been presented so that applications of the SASR can be extended to more general fields. This extension of the SASR method has been adopted to investigate the relationship between nucleosome formation and the ~10bp periodicity of dinucleotides. In general, the result supports the "sequence-specific" argument of nucleosome formation. It suggests that some dinucleotides may connect their ~10bp periodicity with nucleosome formation only in some local regions.

Meanwhile, it implies that the ~10bp periodicity of the sequence is related to not only the occurrence of nucleosome formation, but also the binding preference for the phase in the ~10bp period. Besides, an improved SASR named “mature SASR” has been developed in this chapter to make TP walks of random sequences unbiasedly move around the zero point. Experiments on simulated datasets show the ability of the mature SASR in detecting a hypothetical anti-TP property in DNA sequences. Some real DNA fragments have been found with such an anti-TP property by using the mature SASR. However, the universality of this property in genomes and its biological interpretation are currently unknown.

CHAPTER 6

CONCLUSIONS AND PROSPECTS

6.1 Achievements and limitations

In order to extract and represent periodicity properties hidden in DNA sequences (as objective 1 proposed in Chapter 1), a Self Adaptive Spectral Rotation (SASR) approach has been proposed. At each position in a given DNA sequence, a special vector is calculated from the posterior subsequence. After that, from the generated sequence of the vectors, a random walk in the complex plane is obtained as the SASR's graphic output. It is shown theoretically and experimentally that walk patterns in this graphic output reveal locations of periodic regions as well as the phase shifts (frame shifts) between them: fast moving patterns indicate periodic regions, stable points indicate non-periodic regions, and corners' shapes reveal phase shifts.

The SASR is originally developed for protein-coding region prediction (as objective 2 proposed in Chapter 1), in which the Triplet Periodicity (TP) property is concerned. The tests on real genomic datasets show that the graphic patterns for coding and non-coding regions differ to a great extent, so that the coding regions can be visually distinguished from the non-coding ones. Compared with some other TP-based methods, e.g. the OSCM and the SRM, the SASR has an advantage of higher

sensitivity and specificity in discriminating coding regions. Compared with some currently popular methods, especially the HMM-based methods, the SASR also has some significant advantages including:

1) The SASR does not require any preceding training process, so it can work before any extra information is available, especially helpful when dealing with new sequences from unknown organisms;

2) Without a fixed analysis scale, the graphic output provides opportunities to analyze sequences in various scales and takes advantages of "auto-scale analysis ability" from human vision.

However, as a visualization method, the SASR does not provide a computational result of the prediction. Therefore, two computational approaches, i.e., the FSND and the T-Z-T, have been developed to extract numerical information from the graphic result of the SASR. Among them, the T-Z-T, which consists of two *t*-test segmentation steps and a *z*-test filter step, is considered to be a better and recommended method in this study. The combination of the SASR and the T-Z-T (denoted as SASR-TZT) provides computational predictions of coding regions, maintaining the main advantages of the SASR: The SASR-TZT does not require any training process, either. Meanwhile, the analysis scale is flexibly controlled by three significance levels. Moreover, an experimental comparison between the SASR-TZT and some HMM-based methods shows that small errors in the input DNA sequence

do not greatly influence the performances of this SASR-based approach, so the prediction is more robust.

As a limitation, the newly proposed prediction approach is less accurate than those most popular HMM-based methods when those models are well trained. It is due to two main reasons: First, the SASR-TZT detects sequence regions with high local persistency of TP profile, assuming that these regions match true coding regions to some extent and can be considered as coding region candidates. When the relationship between genetic coding and the TP property becomes weak, in some regions of the DNA sequence, the accuracy may reduce. Second, as its advantage, the graphic output provides opportunities to analyze sequences in various scales, but the visualization is a "fuzzy" method, in which it is difficult to determine the exact boundaries of regions either manually or computationally. So it introduces errors in predicting the boundaries, especially reduces the accuracy in dealing with very short regions.

Considering both the advantages and the limitations mentioned above, the newly proposed prediction method is deemed as an efficient tool, especially for the early stage study on a newly sequenced DNA, when there is insufficient training set and even when the input data is inaccurate. So, objective 2 proposed in Chapter 1 has been achieved.

The SASR approach has been extended to investigate the relationship between

nucleosome formation and the ~10bp periodicity property of dinucleotides in DNA sequences (as objective 3 proposed in Chapter 1). The data of nucleosome formation in *C. elegans* chromosomes have been collected from public databases and their profile functions can be generated (functions *f* and *h*, for occurrence of nucleosome formation and the phase-preferred nucleosome formation respectively). By using the SASR, a "Rightward Rate measure (RR)" spectrum can be generated for any combination of the 16 dinucleotides, representing the intensity of the local ~10bp periodicity along a DNA sequence. A Genetic Algorithm (GA) based method has been developed to find out the dinucleotide combination, which makes the RR spectrum most correlated to the profile functions.

As a result, objective 3 proposed in Chapter 1 has been achieved: The experiment shows that the ~10bp periodicity in dinucleotide sequences are significantly related to nucleosome formation, supporting the argument that nucleosome formation is sequence-specific. The ~10bp periodicity of dinucleotide "TA" has been found to be related to the phase-preferred nucleosome formation in the most extensive places of chromosomes. This result conforms to the finding from Takasuka and Stein (2010). Meanwhile, instead of aligning the nucleosomes to center or beginning, the comparison between nucleosome profile functions and RR spectrums along a DNA sequence allows for the revelation of a less extensive local relationship between dinucleotide periodicity and nucleosome formation. The

experiment suggests that the ~10bp periodicity of some dinucleotides may be related to nucleosome formation only in some local regions. Furthermore, the comparison between the applications using different profile functions h and f implies that the ~10bp periodicity in dinucleotide sequences is related to not only the occurrence of nucleosome formation, but also the binding preference for the phase in the ~10bp period.

By investigating the numerical relationship between the data of nucleosome formation and sequence periodicity, the above experiment supports the "sequence-specific argument" only statistically. It lacks investigation into the intrinsic mechanism of nucleosome formation. Therefore, it does not provide a complete explanation for the principle of nucleosome formation and the debates are still open.

By discussing the essence of the TP property (a persistent pattern) and the randomness, a hypothetical anti-TP property has been presented as an anti-persistent pattern probably contained in DNA sequences. Another extension of the SASR, i.e., the mature SASR, has been adopted to visualize such an anti-TP property in DNA sequences. The application to simulated datasets reveals a high ability of the mature SASR in discriminating the potential anti-TP property, compared with those of other methods. Meanwhile, by applying the mature SASR to real DNA sequences, some real DNA fragments have been found with high intensity of the anti-TP property. It implies that the anti-TP patterns may carry some information. However, this study

only focuses on the ability of the mature SASR in detecting such a potential property. The universality of this property in genomes and its biological interpretation are not known yet.

6.2 Prospects for future work

For the application of the SASR-TZT approach to coding region prediction, one of the main limitations is its accuracy in dealing with region boundaries. For this reason, it is expected to combine the SASR-based method with some other methods sensitive to boundary signals. However, as a tradeoff, such a combination might degenerate the "training-free" feature and the robustness of the original SASR. Such a tradeoff should be concerned under different situations of applications. Meanwhile, as mentioned in the above section, the graphic output of the SASR provides opportunities to analyze sequences in various scales. Although the T-Z-T method controls the analysis scale in a flexible way, it degenerates the "scale-free" feature of the SASR itself and does not take full advantages of "auto-scale analysis ability" from human vision. A possible solution to this problem might be to develop a "semi-automatic" program, combining both manual and computational analyses, in which computer-aided manual analysis could be conducted to roughly and fuzzily partition the graphic output of the SASR and computational methods, such as the T-Z-T, are applied to refine the results.

For the research on nucleosome formation, although our study, to some extent, supports the “sequence-specific” argument and shows the utility of the SASR-based method for this issue, only studying the numerical relationship between nucleosome data and sequence periodicity is not enough to end the debates. With the hints from numerical studies, an investigation into the intrinsic mechanism is expected in future, so that the principle of nucleosome formation can be revealed more clearly and solidly.

Lastly, the anti-TP (anti-persistent) is a hypothetical property newly proposed in this study, opposite to the TP (persistent). Although the mature SASR is capable of detecting the anti-TP property in DNA sequences, it is in doubt whether such a property comprehensively exists in genomes and whether the real DNA fragments found with the anti-TP property do carry any biological information. Since the biological interpretation of this special sequence pattern remains unknown, substantial work should be conducted for this issue in future.

REFERENCES

Adams M.D., Kelley J.M., Gocayne J.D. et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*. **252**(5013). pp. 1651-1656.

Akhtar M., Ambikairajah E. and Epps J. (2008a). Optimizing period-3 methods for eukaryotic gene prediction. In: *Processing of IEEE International Conference on Acoustics, Speech and Signal*. pp. 621-624.

Akhtar M., Epps J. and Ambikairajah E. (2008b). Signal processing in sequence analysis: advances in eukaryotic gene prediction. *IEEE Journal of Selected Topics in Signal Processing*. **2**(3). pp. 310-321.

Alain F.C. and Floyd V.M. (1993). *Gregor Mendel's Experiments on plant hybrids: a guided study*. Rutgers University Press, New Brunswick, US.

Albert I., Mavrich T.N., Tomsho L.P., Qi J., Zanton S.J., Schuster S.C. and Pugh B.F. (2007). Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature*. **446**(7135). pp. 572-576.

Alberts B., Bray D., Lewis J., Raff M., Robert K. and Watson J.D. (1994). *Molecular Biology of the Cell, Third Edition.* Garland Publishing, New York, US.

Almeida J.S., Carriço J.A., Maretzek A., Noble P.A. and Fletcher M. (2001). Analysis of genomic sequences by chaos game representation. *Bioinformatics*. **17**(5). pp. 429-437.

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. (1990). Basic local

alignment search tool. *Journal of Molecular Biology*. **215**(3). pp. 403-410.

Anastassiou D. (2000). Frequency-domain analysis of biomolecular sequences. *Bioinformatics*. **16**(12). pp. 1073-1081.

Anastassiou D. (2001). Genomic signal processing. *Bioinformatics. Signal Processing Magazine*. **18**(4). pp. 8-20.

Aparicio O., Geisberg J.V. and Struhl K. (2004). Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. In: *Current Protocols in Cell Biology*. John Wiley, New York, US.

Avery O.T., MacLeod C.M. and McCarty M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus Type III. *Journal of Experimental Medicine*. **79**(2). pp. 137-158.

Baldi P. and Brunak S. (1998). *Bioinformatics: the machine learning approach.* MIT Press, Cambridge, UK.

Barciszewski J. and Erdmann V.A. (2003). *Noncoding RNAs: molecular biology and molecular medicine*. Springer, New York, US.

Barski A., Cuddapah S., Cui K., Roh T.Y., Schones D.E., Wang Z., Wei G., Chepelev I. and Zhao K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*. **129**(4). pp. 823-837.

Beadle G.W. and Tatum E.L. (1941). Genetic control of biochemical reactions in Neurospora. *Proceedings of the National Academy of Sciences USA*. **27**(11). pp. 499-506.

Bennetzen J.L. and Hall B.D. (1982). Codon selection in yeast. *Journal of Biological Chemistry*. **257**(6). pp. 3026-3031.

Berget S.M., Moore C. and Sharp P.A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences USA*. **74**(8). pp. 3171-3175.

Berg J.M., Tymoczko J.L. and Stryer L. (2002). *Biochemistry (5th edition)*. W.H. Freeman, New York, US.

Bernstein B.E., Liu C.L., Humphrey E.L., Perlstein, E.O. and Schreiber S.L. (2004). Global nucleosome occupancy in yeast. *Genome Biology*. **5**(9). R62.

Bernaola-Galvan P., Ivanov P. Ch., Nunes Amaral L. A. and Stanley H. E. (2001). Scale invariance in the nonstationarity of human heart rate. *Physical Review Letters*. **87**(16). doi:10.1103/PhysRevLett.87.168105.

Berthelsen C.L., Glazier J.A. and Skolnick M.H. (1992). Global fractal dimension of human DNA sequences treated as pseudorandom walks. *Physical Review A*. **45**(12). pp. 8902-8913.

Birney E. and Durbin R. (2000). Using GeneWise in the Drosophila annotation experiment. *Genome Research*. **10**(4). pp. 547-548.

Borodovsky M. and McIninch J. (1993). GeneMark: Parallel gene recognition for both DNA strands. *Computers & Chemistry*. **17**(2). pp. 123-133.

Burge C. and Karlin S. (1997). Prediction of complete gene structures in Human genomic DNA. *Journal of Molecular Biology*. **268**(1). pp. 78-94.

Burge C.B. and Karlin S. (1998). Finding the genes in genomic DNA. *Current Opinion in Structure Biology*. **8**(3). pp. 346-354.

Bussemakers M.J.G., Van Bokhoven A., Verhaegh G.W., Smit F.P., Karthaus H.F.M., Schalken J.A., Debruyne F.M.J., Ru N. and Isaacs W.B. (1999). DD3: A new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Research*. **59**(23). pp. 5975-5979.

Cairns B.R. (2005). Chromatin remodeling complexes: strength in diversity, precision through specialization. *Current Opinion in Genetics & Development*. **15**(2). pp. 185-190.

Cao Y.H., Tung W.W., Gao J.B. and Qi Y. (2005). Recurrence time statistics: Versatile tools for genomic DNA sequence analysis. *Journal of Bioinformatics and Computational Biology*. **3**(3). pp. 677-696.

Cebrat S. and Dudek M.R. (1996). Generation of overlapping open reading frames. *Trends in Genetics*. **12**(1). pp. 12.

Cebrat S. and Dudek M.R. (1998). The effect of DNA phase structure on DNA walks. *The European Physical Journal B*. **3**(2). pp. 271-276.

Chang C.Q., Fung Peter C.W. and Hung Y.S. (2008). Improved gene prediction by resampling-based spectral analysis of DNA sequence. In: *Proceedings of the 5th International Conference on Information Technology and Application in Biomedicine, in conjunction with The 2nd International Symposium & Summer School on Biomedical and Health Engineering*. Shenzhen, China. pp. 221-224.

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

Chargaff E., Lipshitz R., Green C. and Hodes M.E. (1951). The composition of the

deoxyribonucleic acid of salmon sperm. *Journal of Biological Chemistry.* **192**(1). pp. 223-230.

Chargaff E., Lipshitz R. and Green C. (1952). Composition of the deoxypentose nucleic acids of four genera of sea-urchin. *Journal of Biological Chemistry.* **195**(1). pp. 155-160.

Chen K., Meng Q., Ma L., Liu Q., Tang P., Chiu C., Hu S. and Yu J. (2008). A novel DNA sequence periodicity decodes nucleosome positioning. *Nucleic Acids Research*. **36**(19). pp. 6228-6236.

Chen K., Wang L., Yang M., Liu J., Xin C., Hu S. and Yu J. (2010) Sequence signatures of nucleosome positioning in Caenorhabditis elegans. *Genomics, Proteomics & Bioinformatics*. **8**(2). pp. 92-102.

Claverie J.M. and Bougueleret L. (1986). Heuristic informational analysis of sequences. *Nucleic Acids Research*. **14**(1). pp. 179-196.

Clegg R.G. (2006). A practical guide to measuring the Hurst parameter. *International Journal of Simulation: Systems, Science & Technology.* **7**(2). pp. 3-14.

Cooper G.M. (2000). *The Cell: A Molecular Approach (2nd edition)*. Sinauer Associates, Sunderland, UK.

Costantini M. and Bernardi G. (2008). The short-sequence designs of isochores from the human genome. *Proceedings of the National Academy of Sciences USA*. **105**(37). pp. 13971-13976.

Crick F.H.C. (1958). Ideas on protein synthesis. *Symposia of the Society for Experimental Biology.* **XII**. pp. 139-163.

Crothers D.M., Drak J., Kahn J.D. and Levene S.D. (1992). DNA bending, flexibility,

and helical repeat by cyclization kinetics. *Methods in Enzymology*. **212**. pp. 3-29.

Dahm R. (2008). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics.* **122**(6). pp. 565-581.

Datta S. and Asif A. (2004). DFT based DNA splicing algorithms for prediction of protein coding regions. In: *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*. pp. 45-49.

Davey C., Pennings S., Meersseman G., Wess T.J. and Allan J. (1995). Periodicity of strong nucleosome positioning sites around the chicken adult beta-globin gene may encode regularly spaced chromatin. *Proceedings of the National Academy of Sciences USA*. **92**(24). pp. 11210-11214.

Dayhoff M.O., Schwartz R.M. and Orcutt. B.C. (1978). *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, US.

De Santis P., Palleschi A., Savino M. and Scipioni A. (1988). A theoretical model of DNA curvature. *Biophysical Chemistry*. **32**(2-3). pp. 305-317.

De Santis P., Palleschi A., Savino M. and Scipioni A. (1990). Validity of the nearest-neighbor approximation in the evaluation of the electrophoretic manifestations of DNA curvature. *Biochemistry*. **29**(39). pp. 9269-9273.

Deschavanne P.J. (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular Biology and Evolution*. **16**(10). pp. 1391-1399.

De Gennes P.G. (1999). Passive entry of a DNA molecule into a small pore. *Proceedings of the National Academy of Sciences USA*. **96**(13). pp. 7262-

7264.

Devynck P., Wang G., Antar G. and Bonhomme G. (2000). The Hurst exponent and long-time correlation. In: *Proceedings of EPS 27th Conference on Controlled Fusion and Plasma Physics.* Budapest, Hungary. pp. 12-16.

Dlakic M., Ussery D.W. and Brunak S. (2005). DNA bendability and nucleosome positioning in transcriptional regulation. In: *DNA conformation and transcription*. Springer, New York, US. pp. 189-202.

Dodin G., Levoir P. and Cordier C. (1996). Triplet correlation in DNA sequences and stability of heteroduplexes. *Journal of Theoretical Biology*. **183**(3). pp. 341-343.

Dodin G., Vandergheynst P., Levoir P., Cordier C. and Marcourt L. (2000). Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. *Journal of Theoretical Biology*. **206**(3). pp. 323-326.

Doherty E.A. and Doudna J.A. (2001). Ribozyme structures and mechanisms. *Annual Review of Biophysics and Biomolecular Structure*. **30**. pp. 457-475.

Do J.H. and Choi D.K. (2006). Computational approaches to gene prediction. *Journal of Microbiology*. **44**(2). pp. 137-144.

Dong S. and Searls D.B. (1994). Gene structure prediction by linguistic methods. *Genomics*. **23**(3). pp. 540-551.

Doudna J.A. and Cech T.R. (2002). The chemical repertoire of natural ribozymes. *Nature*. **418**(6894). pp. 222-228.

Eftestol T., Ryen T., Aase S.O., Strassle C., Boos M., Schuster G. and Ruoff P. (2006). Eukaryotic gene prediction by spectral analysis and pattern recognition

techniques. In: *Proceedings of 7th Nordic Signal Processing Symposium*. Reykjavik, Iceland. pp. 146-149.

Ephraim Y. and Merhav N. (2002). Hidden Markov processes. *IEEE Transactions on Information Theory*. **48**(6). pp. 1518-1569.

Epps J., Ambikairajah E. and Akhtar M. (2008). An integer period DFT for biological sequence processing. In: *Proceedings of IEEE International Workshop on Genomic Signal Processing and Statistics*. Phoenix, US. pp. 1-4.

Feder J. (1988). *Fractals*. Plenum Press, New York, US.

Fickett J.W. (1982). Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research*. **10**(17). pp. 5303-5318.

Fickett J.W. and Tung C.S. (1992). Assessment of protein coding measures. *Nucleic Acids Research*. **20**(24). pp. 6441-6450.

Fickett J.W. (1996). The gene identification problem: An overview for developers. *Computers and Chemistry*. **20**(1). pp. 103-118.

Frank D.N. and Pace N.R. (1998). Ribonuclease P: unity and diversity in a tRNA processing ribozyme. *Annual Review of Biochemistry*. **67**. pp. 153-180.

Franklin R. and Gosling R. (1953). Molecular configuration in Sodium Thymonucleate. *Nature*. **171**(4356). pp. 740-741.

Frenkel F.E. and Korotkov E.V. (2008). Classification analysis of triplet periodicity in protein-coding regions of genes. *Gene*. **421**(1-2). pp. 52-60.

Frenkel F.E. and Korotkov E.V. (2009). Using triplet periodicity of nucleotide sequences for finding potential reading frame shifts in genes. *DNA Research*. **16**(2). pp. 105-114.

Fuentes A.R., Ginori J.V.L. and Ábalo R.G. (2006). Detection of coding regions in large DNA sequences using the short time Fourier transform with reduced computational load. In: *Progress in Pattern Recognition, Image Analysis and Applications*. pp. 902-909.

Gao J.B., Qi Y., Cao Y.H. and Tung W.W. (2005). Protein coding sequence identification by simultaneously characterizing the periodic and random features of DNA sequences. *Journal of Biomedicine and Biotechnology*. **2005**(2). pp. 139-146.

Gelfand M.S., Mironov A.A. and Pevzner P.A. (1996). Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences USA*. **93**(17). pp. 9061-9066.

Gerstein M.B., Bruce C., Rozowsky J.S., Zheng D.Y., Du J., Korbel J.O., Emanuelsson O., Zhang Z.D. D., Weissman S. and Snyder M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research.* **17**(6). pp. 669-681.

Ghaemmaghami S., Huh W.K., Bower K., Howson R.W., Belle A., Dephoure N., O'Shea E.K. and Weissman J.S. (2003). Global analysis of protein expression in yeast. *Nature*. **425**(6959). pp. 737-741.

Goad W.B. and Kanehisa M.I. (1982). Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nucleic Acids Research*. **10**(1). pp. 247-263.

Griffith F. (1928). The significance of pneumococcal types. *The Journal of Hygiene.* **27**(2). pp. 113-159.

Guigó R., Agarwal P., Abril J.F., Burset M., Fickett J.W. (2000). An assessment of gene prediction accuracy in large DNA sequences. *Genome Research*. **10**(10). pp. 1631-1642.

Guigó R., Knudsen S., Drake N. and Smith T. (1992). Prediction of gene structure. *Journal of Molecular Biology*. **226**(1). pp. 141-157.

Guillemette B., Bataille A.R., Gévry N., Adam M., Blanchette M., Robert F. and Gaudreau L. (2005). Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. *PLoS Biology*. **3**(12). e384. doi:10.1371/journal.pbio.0030384.

Guo F.B., Ou H.Y. and Zhang C.T. (2003). ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Research*. **31**(6). pp. 1780-1789.

Haimovich A.D., Byrne B., Ramaswamy R. and Welsh W.J. (2006). Wavelet analysis of DNA walks. *Journal of Computational Biology*. **13**(7). pp. 1289-1298.

Hamilton J. D. (2008). Regime-switching models. In: *The New Palgrave Dictionary of Economics (2nd edition)*. Palgrave McMillan Ltd, Basingstoke, UK.

Hosid S., Trifonov E.N. and Bolshoy A. (2004). Sequence periodicity of Escherichia coli is concentrated in intergenic regions. *BMC Molecular Biology*. **5**(14). doi:10.1186/1471-2199-5-14.

Huang X., Ariki Y. and Jack M. (1990). *Hidden Markov Models for Speech Recognition*. Columbia University Press New York, New York, US.

Jeffrey H. (1990). Chaos game representation of gene structure. *Nucleic Acids Research*. **18**(8). pp. 2163-2170.

Jenuwein T. and Allis C.D. (2001). Translating the histone code. *Science*. **293**(5532). pp. 1074-1080.

Jiang C. and Pugh B.F. (2009). Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics*. **10**(3). pp. 161-172.

Jiang X.Y., Lavenier D. and Yau Stephen S.T. (2008). Coding region prediction based on a universal DNA sequence representation method. *Journal of Computational Biology*. **15**(10). pp. 1237-1256.

Johnson S.M., Tan F.J., McCullough H.L., Riordan D.P. and Fire A.Z. (2006) Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Research*. **16**(12). pp. 1505-1516.

Kamakaka R.T. and Biggins S. (2005). Histone variants: deviants? *Genes and Development*. **19**(3). pp. 295-310.

Kaplan N., Moore I.K., Fondufe-Mittendorf Y., Gossett A.J., Tillo D., Field Y., LeProust E.M., Hughes T.R., Lieb J.D., Widom J. and Segal E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*. **458**(7236). pp. 362-366.

Kasamatsu H., Robberson D.L. and Vinograd J. (1971). A novel closed-circular mitochondrial DNA with properties of a replicating intermediate. *Proceedings of the National Academy of Sciences USA*. **68**(9). pp. 2252-2257.

Kiyama R. and Trifonov E.N. (2002). What positions nucleosomes? -- A model. *FEBS Letters*. **523**(1). pp. 7-11.

Kornberg R.D. and Lorch Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*. **98**(3). pp. 285-294.

Korotkov E.V., Korotkova M.A., Frenkel F.E. and Kudryashov N.A. (2003). The informational concept of searching for periodicity in symbol sequences. *Molecular Biology*. **37**(3). pp. 372-386.

Kotlar D. and Lavner Y. (2003). Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. *Genome Research*. **13**(8). pp. 1930-1937.

Krogh A. (1997). Two methods for improving performance of an HMM and their application for gene finding. In: *Proceedings of International Conference on Intelligent Systems for Molecular Biology*. **5**. pp.179-186.

Kulp D., Haussler D., Reese M.G. and Eeckman F.H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. In: *Proceedings of International Conference on Intelligent Systems for Molecular Biology*. **4**. pp. 134-142.

Lee C.K., Shibata Y., Rao B., Strahl B.D. and Lieb J.D. (2004). Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics*. **36**(8). pp. 900-905.

Lee W., Tillo D., Bray N., Morse R.H., Davis R.W., Hughes T.R. and Nislow C. (2007). A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genetics*. **39**(10). pp. 1235-1244.

Levene P. (1919). The structure of yeast nucleic acid. *Journal of Biological*

Chemistry. **40**(2). pp. 415-424.

Levene S.D., Wu H.M. and Crothers D.M. (1986). Bending and flexibility of kinetoplast DNA. *Biochemistry*. **25**(14). pp. 3988-3995.

Liu A.Y., Torchia B.S., Migeon B.R. and Siliciano R.F. (1997). The human NTT gene: identification of a novel 17-kb noncoding nuclear RNA expressed in activated CD4+ T cells. *Genomics*. **39**(2). pp. 171-184.

Li W. (1997). The complexity of DNA. *Complexity*. **3**(2). pp. 33-37.

Lowary P.T. and Widom J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal Molecular Biology*. **276**(1), 19-42.

Luger K., Mader A.W., Richmond R.K., Sargent D.F. and Richmond T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*. **389**(6648). pp. 251-260.

Mandelbrot B.B. and Hudson R.L. (2004). *The (mis) behavior of market: A fractal view of risk, ruin, and reward.* Basic Book, New York, US.

Marini J.C., Levene S.D., Crothers D.M. and Englund P.T. (1982). Bent helical structure in kinetoplast DNA. *Proceedings of the National Academy of Sciences USA*. **79**(24). pp. 7664-7668.

Masoom H., Datta S., Asif A., Cunningham L. and Wu G. (2006). A fast algorithm for detecting frame shifts in DNA sequences. In: *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. pp. 1-8.

Mattick J.S. (1994). Introns: evolution and function. *Current Opinion in Genetics*

and Development. **4**(6). pp. 823-831.

Mavrich T.N., Jiang C., Ioshikhes I.P., Li X., Venters B.J., Zanton S.J., Tomsho L.P., Qi J., Glaser R.L., Schuster S.C., Gilmour D.S., Albert I. and Pugh B.F. (2008). Nucleosome organization in the Drosophila genome. *Nature*. **453**(7193). pp. 358-362.

Nair T.M., Madhusudan K., Nagaraja V., Kulkarni B.D., Majumdar H.K. and Singh R. (1994). On the mobility behavior of a curved DNA fragment located in circular permutation. *FEBS Letters*. **351**(3). pp. 321-324.

Nair T.M. (2009). Sequence periodicity in nucleosomal DNA and intrinsic curvature. *BMC Structural Biology*. **10**(Suppl 1). S8. doi:10.1186/1472-6807-10-S1-S8.

Neidle S. (2008). *Principles of Nucleic Acid Structure.* Academic Press.

Oliver J.L. and Bernaola P. (1993). Entropic profiles of DNA sequences through chaos-game-derived images. *Journal of Theoretical Biology.* **160**(4). pp. 457-470.

Olson D.L. and Delen D. (2008). *Advanced Data Mining Techniques*. Springer-Verlag Berlin Heidelberg.

Orlov Y.L., Te Boekhorst R. and Abnizova I. (2006). Statistical measures of the structure of genomic sequences: entropy, complexity and position information. *Journal of Bioinformatics and Computational Biology*. **4**(2). pp. 523-526.

Pearson K. (1905). The problem of the random walk. *Nature*. **72**. pp. 294, 318, 342.

Peng C.K., Buldyrev S.V., Goldberger A.L., Havlin S., Sciortino F., Simons M. and

Stanley H.E. (1992). Long-range correlations in nucleotide sequences. *Nature*. **356**(6365). pp. 168-170.

Rabiner L.R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of IEEE*. **77**(2). pp. 257-286.

Randić M., Zupan J. and Balaban A.T. (2004). Unique graphical representation of protein sequences based on nucleotide triplet codons. *Chemical Physics Letters*. **397**(1-3). pp. 247-252.

Randić M. (2008). Another look at the chaos-game representation of DNA. *Chemical Physics Letters*. **456**(1-3). pp. 84-88.

Reese M.G., Eeckman F.H., Kulp D. and Haussler D. (1997). Improved splice site detection in Genie. *Journal of Computational Biology*. **4**(3). pp. 311-323.

Ré M. and Pavesi G. (2009). Detecting conserved coding genomic regions through signal processing of nucleotide substitution patterns. *Artificial Intelligence in Medicine*. **45**(2). pp. 117-123.

Richmond T.J. and Davey C.A. (2003). The structure of DNA in the nucleosome core. *Nature*. **423**(6936). pp. 145-150.

Ricker D.W. (2003). *Echo Signal Processing*. Kluwer Academic Publishers, Dordrecht, Netherlands.

Salih F., Salih B., Trifonov E.N. (2008). Sequence structure of hidden 10.4-base repeat in the nucleosomes of C. elegans. *Journal of Biomolecular Structure and Dynamics*. **26**(3). pp. 273-282.

Salzberg S.L., Delcher A.L., Kasif S. and White O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*.

26(2). pp. 544-548.

Sarma K. and Reinberg D. (2005). Histone variants meet their match. *Nature Reviews. Molecular Cell Biology*. **6**(2). pp. 139-149.

Satchwell S.C., Drew H.R. and Travers A.A. (1986). Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology*. **191**(4). pp. 659-675.

Schwabish M.A. and Struhl K. (2004). Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II. *Molecular Cellular Biology*. **24**(23). pp. 10111-10117.

Segal E., Fondufe-Mittendorf Y., Chen L., Thåström A., Field Y., Moore I.K., Wang J.P. and Widom J. (2006). A genomic code for nucleosome positioning. *Nature*. **442**(7104). pp. 772-778.

Segal E. and Widom J. (2009). Poly(dA:dT) tracts: major determinants of nucleosome organization. *Current Opinion in Structural Biology*. **19**(1). pp. 65-71.

Sekinger E.A., Moqtaderi Z. and Struhl K. (2005). Intrinsic histone–DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Molecular Cell*. **18**(6). pp. 735-748.

Sellers P.H. (1984). Pattern recognition in genetic sequences by mismatch density. *Bulletin of Mathematical Biology*. **46**(4). pp. 501-514.

Shivaswamy S., Bhinge A., Zhao Y., Jones S., Hirst M. and Iyer V.R. (2008). Dynamic remodeling of individual nucleosomes across a eukaryotic

genome in response to transcriptional perturbation. *PLoS Biology*. **6**(3). e65.

Smith T.F. and Waterman M.S. (1981). Comparison of biosequences. *Advances in Applied Mathematics*. **2**(4). pp. 482-489.

Snyder E.E. and Stormo G.D. (1993). Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Research*. **21**(3). pp. 607-613.

Snyder E.E. and Stormo G.D. (1995). Identification of protein coding regions in genomic DNA. *Journal Molecular Biology*. **248**(1). pp. 1-18.

Solovyev V.V., Salamov A.A. and Lawrence C.B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Research*. **22**(24). pp. 5156-5163.

Staden R. and McLachlan A.D. (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Research*. **10**(1). pp. 141-156.

Stanke M. and Waack S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*. **19**(2). pp. ii215-ii225.

Stewart I. (2001). Where drunkards hang out. *Nature*. **413**(6857). pp. 686-687.

Stormo G.D. (2000). Gene-finding approaches for eukaryotes. *Genome Research*. **10**(4). pp. 394-397.

Takasuka T.E. and Stein A. (2010). Direct measurements of the nucleosome-forming preferences of periodic DNA motifs challenge established models. *Nucleic Acids Research*. **38**(17). pp. 5672-5680.

Te Boekhorst R., Abnizova I. and Nehaniv C. (2008). Discriminating coding, non-coding and regulatory regions using rescaled range and detrended fluctuation analysis. *BioSystems*. **91**(1). pp. 183-194.

Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S. and Ramaswamy R. (1997). Prediction of probable genes by Fourier analysis of genomic sequences. *Computer Applications in the Biosciences*. **13**(3). pp. 263-270.

Trifonov E.N. (1980). Sequence-dependent deformational anisotropy of chromatin DNA. *Nucleic Acids Research*. **8**(17). pp. 4041-4053.

Trifonov E.N. and Sussman J.L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proceedings of the National Academy of Sciences USA*. **77**(7). pp. 3816-3820.

Trifonov E.N. (2010). Base pair stacking in nucleosome DNA and bendability sequence pattern. *Journal of Theoretical Biology*. **263**(3). pp. 337-339.

Tuqan J. and Rushdi A. (2006). The filtered spectral rotation measure. In: *Proceedings of 40th Asilomar Conference on Signals, Systems and Computers*. pp. 1875-1879.

Tuqan J. and Rushdi A. (2008). A DSP approach for finding the codon bias in DNA sequences. *IEEE Journal of Selected Topics in Signal Processing*. **2**(3). pp. 343-356.

Uberbacher E.C. and Mural R.J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proceedings of the National Academy of Sciences USA*. **88**(24). pp. 11261-11265.

Ulanovsky L.E. and Trifonov E.N. (1987). Estimation of wedge components in curved DNA. *Nature*. **326**(6114). pp. 720-722.

Vaglica G., Lillo F., Moro E. and Mantegna R. N. (2008). Scaling laws of strategic behavior and size heterogeneity in agent dynamics. *Physical Review E*. **77**(3). doi:10.1103/PhysRevE.77.036110.

Valouev A., Ichikawa J., Tonthat T., Stuart J., Ranade S., Peckham H., Zeng K., Malek J.A., Costa G., McKernan K., Sidow A., Fire A. and Johnson S.M. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research*. **18**(7). pp. 1051-1063.

Van Holde K.E. (1989) *Chromatin*. Springer, New York, US.

Voss R.F. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical Review Letters.* **68**(25). pp. 3805-3808.

Wain-Hobson S. (2006). The third bond. *Nature*. **439**(7076). pp. 539.

Wang Y., Crawford D.R. and Davies K.J. (1996). Adapt33, a novel oxidant-inducible RNA from hamster HA-1 cells. *Archives of Biochemistry and Biophysics*. **332**(2). pp. 255-260.

Wang Y., Hill K., Singh S. and Kari L. (2005). The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene*. **346**. pp. 173-185.

Wang Z., Chen Y. and Li Y. (2004). A brief review of computational gene prediction methods. *Genomics, Proteomics & Bioinformatics*. **2**(4). pp. 216-221.

Watson J.D. and Crick F.H.C. (1953). A structure for deoxyribose nucleic acid. *Nature*. **171**(4356). pp. 737-738.

Widom J. (1996). Short-range order in two eukaryotic genomes: Relation to chromosome structure. *Journal of Molecular Biology*. **259**(4). pp. 579-588.

Widom J. (2001). Role of DNA sequence in nucleosome stability and dynamics. *Quarterly Reviews of Biophysics*. **34**(3). pp. 269-324.

Wu H.M. and Crothers D.M. (1984). The locus of sequence-directed and protein-induced DNA bending. *Nature*. **308**(5959). pp. 509-513.

Wyrick J.J., Holstege F.C., Jennings E.G., Causton H.C., Shore D., Grunstein M., Lander E.S. and Young R.A. (1999). Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature*. **402**(6760). pp. 418-421.

Yan M., Lin Z.S. and Zhang C.T. (1998). A new Fourier transform approach for protein coding measure based on the format of the Z curve. *Bioinformatics*. **14**(8). pp. 685-690.

Yin C.C. and Yau Stephen S.T. (2007). Prediction of protein coding regions by 3-base periodicity analysis of a DNA sequence. *Journal of Theoretical Biology*. **247**(4). pp. 687-694.

Yuan G.C., Liu Y.J., Dion M.F., Slack M.D., Wu L.F., Altschuler S.J. and Rando O.J. (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*. **309**(5734). pp. 626-630.

Yu Z.G., Anh V. and Lau K.S. (2004). Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. *Journal of Theoretical Biology.* **226**(3). pp. 341-348.

Zanton S.J. and Pugh B.F. (2006). Full and partial genomewide assembly and

disassembly of the yeast transcription machinery in response to heat shock. *Genes and Development*. **20**(16). pp. 2250-2265.

Zhang C.T. and Wang J. (2000). Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on Z curve. *Nucleic Acids Research*. **28**(14). pp. 2804-2814.

Zhang C.T., Zhang R. and Ou H.Y. (2003). The Z curve database: a graphic representation of genome sequences. *Bioinformatics*. **19**(5). pp. 593-599.

Zhang M.Q. (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proceedings of the National Academy of Sciences USA*. **94**(2). pp. 565-568.

Zhang R. and Zhang C.T. (1994). Z-curves, an intutive tool for visualizing and analyzing the DNA sequences. *Journal of Biomolecular Structure & Dynamics*. **11**(4). pp. 767-782.

Zhang R. and Zhang C.T. (2002). Single replication origin of the archaeon Methanosarcina mazei revealed by the Z curve method. *Biochemical and Biophysical Research Communications.* **297**(2). pp. 396-400.

Zhang R. and Zhang C.T. (2003). Identification of genomic islands in the genome of Bacillus cereus by comparative analysis with Bacillus anthracis. *Physiological Genomics*. **16**(1). pp. 19-23.

Zhang R. and Zhang C.T. (2005). Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea.* **1**(5). pp. 335-346.

Zuccheri G., Scipioni A., Cavaliere V., Gargiulo G., De Santis P. and Samori B. (2001). Mapping the intrinsic curvature and flexibility along the DNA

chain. *Proceedings of the National Academy of Sciences USA*. **98**(6). pp. 3074-3079.