Two Methods of Analyse DNA Sequences:

Predicting Coding Regions and

Clustering Homologous DNA

BY

BO ZHAO
B.A. (Xian Jiaotong University) 1999
M.S. (Xian Jiaotong University) 2002
M.S. (University of Illinois at Chicago) 2006

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathemathics
in the Graduate College of the
University of Illinois at Chicago, 2011

Chicago, Illinois

Defense Committee:

Stephen Yau, Chair and Advisor
Jie Yang
Jan Verschelde
David Nicholls
Wai-Yee Keung, Department of Physics

In the memory of my father, Shuncai Zhao, who passed away at the age of 63 after flighting liver cancer in the summer of 2010. He was still being concerned about my dissertation in his last moment.

R. I. P.

ACKNOWLEDGMENTS

I want to thank Professor Stephen Yau, my thesis advisor. It is impossible for me to achieve the research result and accomplish the dissertation without his successful supervision and support. I also appreciate Professor David Nicholls, Professor Jan Verschelde, Professor Jie Yang and Professor Wai Yee Keung for their unwavering assistance. In addition, I would thank Professor Jiasong Wang from Department of Mathematics of Nanjing Univerisity, China, whose advice and encouragement is important for the thesis.

Moreover, many thanks go to my parents, Shuncai Zhao and Huifang Zhang, and my parents-in-law, Jianhua Hao and Fangying Qiang, for their tremendous supports, both emotionally and occasionally financially.

My greatest thanks goes to my family. My lovely son Steven was born at Chicago one year ago. He really challenges my patience and ability of balancing everything. However, he has shaped my graduate experience more than anything else. My wife, Qiang, who is a Ph.D student at West Virginia University. During the past two years when we live separately, she can take good care of Steven and herself during the pregnancy and infant period, and she did well in her research at the same time. Her constant love and support have made my graduate life a wonderful experience. I am fortunate to have such an amazing parter for life and the eternity.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
NIH	National Institutes of Health
PS	Power Spectrum
TICOR	Threshold to Identify Coding Region
MSA	Multiple Sequence Alignment
DV	Distribution Vectors

SUMMARY

With the exponential growth of DNA sequences in the past twenty years, it has became ineffective to analyze DNA sequences only through the traditional biological experiments. Various mathematical methods and computer algorithms are applied to sequence analyses and related research areas, which help the biological study to be upgraded into automatic programming from manual operation. Especially, there are two important research areas to study DNA sequences in bioinformatics. One is to predict the coding regions on DNA sequences, another is to determine the evolutionary relationship based on DNA sequences. In this thesis, two mathematical methods are introduced to show our achievements in these two research areas respectively.

In chapter two, we introduce a simple parameter called $TICOR$ (Threshold to Identify Coding Region) to distinguish the coding regions from non-coding regions. The method only takes the linear computation time which is much better than those of Fourier Transform and other methods. Moreover, we are able to estimate the proportion of coding regions to the length of the whole DNA sequence simply basing on the parameter $TICOR$. Finally, we develop a novel method to predict the coding regions from DNA sequences, which we call TICORSCAN. We do the test on the ROSETTA dataset(1) with our TICORSCAN method and other popular method, such as GENSCAN(2) and TWINSCAN(3). The prediction accuracy shows that our TICORSCAN method is able to predict the coding regions more efficiently.

Secondly, we report a novel mathematical method to transform the DNA sequences into the

distribution vectors in chapter three. The distribution vectors correspond to points in the sixty dimensional Euclidean space. Each component of the distribution vectors represents the distribution of one kind of nucleotide in k segments of the DNA sequence. The statistical properties of the distribution vectors are demonstrated and examined with huge datasets of human DNA sequences and random sequences. The determined expectation and standard deviation can make the mapping stable and practicable. Moreover, we apply the distribution vectors to the clustering of the mitochondrial complete genomes from 80 placental mammals and the gene Haemagglutinin (HA) of 60 H1N1 viruses from Human, Swine and Avian. The 80 mammals and 60 H1N1 viruses are classified accurately and rapidly compared to the multiple sequence alignment methods. The results indicate that the distribution vectors can reveal the similarity and evolutionary relationship among homologous DNA sequences based on the distances between any two of these distribution vectors. The advantage of fast computation offers the distribution vectors the opportunity to deal with the huge amount of DNA sequences efficiently.

CHAPTER 1

INTRODUCTION

With the maturity of the technology to identify DNA sequences in the last decades, the amount of DNA sequences increases at an incredible speed. For example, Genbank, the National Institutes of Health's (NIH) genetic sequence database, is an annotated collection of all publicly available DNA sequences. There were approximately 117,476,523,128 bases in 122,941,883 sequence records for almost all life-forms in August 2010, whereas there were only 680,338 bases in 606 sequence records in December 1982(4). The number of bases in GenBank has doubled approximately every 18 months since NIH launched the database. In order to analyze the DNA sequences in the huge database, many mathematical methods and computer programs are developed to solve specific problems about analyzing DNA sequences. In this chapter, we will introduce elementary biological background and provide the brief introduction of two important research areas. One is the prediction of Coding Regions in DNA sequences, another is to do the clustering with homologous DNA sequences to discover the evolutionary relationship among the organisms. Later, chapter two describes our TICORSCAN method to predict the coding regions in the DNA sequences. The prediction has the high accuracy compared to other popular methods, such as GENESCAN and TWINSCAN. Furthermore, our distribution vector method will be introduced in the chapter three, which can map the DNA sequences into the Euclidean space and do the clustering. The phylogenetic trees based on the clustering show

that the distances between any two of the distribution vectors correspond to the similarity and evolutionary relationship among these DNA sequences.

1.1 Coding Regions in DNA Sequences

Deoxyribonucleic acid (DNA) is a nucleic acid which carries genetic information for the biological development of all cellular forms of life and many viruses, which consists of two long strands with the double helix structure. Each strand has the direction from 5' end to 3' end and consists four type nucleotides including Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). Adenine pairs with Thymine and Cytosine pairs with Guanine to form the double helix structure.

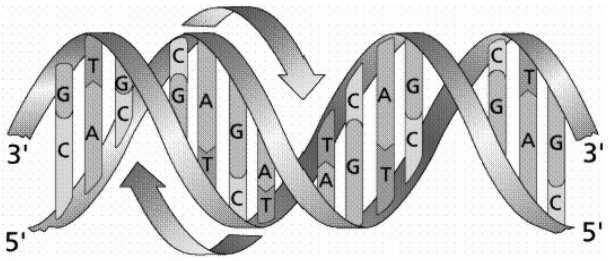

Figure 1. The Double Helix Structure of DNA

The main biological function of DNA sequences is that could be encoded into protein sequences via RNA transcription. However, Not the whole DNA sequences are encoded to protein sequences. Some regions are removed when the DNA sequences are transcribed into RNA, which

we call Introns. The regions transcribed into RNA are called Exons. Exons are encoded into the protein sequences except the untranslated regions, which are important for efficient translation of the transcript and for controlling the rate of translation. We define the coding regions as the part of Exons that are encoded into protein sequences, while we consider the untranslated regions and Introns together as the non-coding regions. Figure 2 explains the partition clearly.

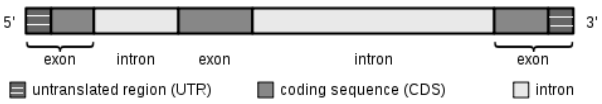

Figure 2. The Partition of Coding Region, Exons and Introns

Traditionally, the exons or coding regions can be found by comparing DNA, RNA and Protein sequences in biological experiment. However, it becomes a challenge to find the Exons or coding regions in DNA sequences with automatic computational methods when the amount of DNA sequences increases rapidly. In 1982, Fickett(5) introduced a statistical method to predict the coding regions regions which is based on simple and universal differences between coding and non-coding DNA sequences. Later, Burset(6) provided the benchmark to evaluate of the prediction programs. Recently, many methods are developed to predict the coding regions in DNA sequences, which includes Dynamic Programming(7), Hidden Markov Model(2), Neural

network(8) and other methods. However, those methods do not show the intrinsic biological difference between coding regions and non-coding regions. They only estimate the parameters of complicated statistical or machine learning models based on the relatively small training data set and apply these models to predict the coding regions. In the chapter two, we introduce our *TICORSCAN* method that can demonstrate the difference between coding regions and non-coding regions by the curve directly. Most importantly, our method can predict the coding regions in the DNA sequences with high accuracy.

1.2 Clustering Homologous DNA Sequences

About one hundred and fifty years ago, Charles Robert Darwin(9) published the famous theory called natural selection in his book On the Origin of Species. He claimed that all species of life in the earth are descended from common ancestors based on the geographical distribution of wildlife and fossils he collected. The idea and a simple draft of the phytogenetic tree was also provided in his book. Later, with the development of evolutionary biology, it necessary to find a new practicable method to construct the phylogenetic tree automatically instead of using fossils manually.

It is well known that DNA is transferred from organisms to their offspring. During the transmission, A few changes always take place in DNA sequences. Hence, the organisms who share a lineage and are descended from a common ancestor must have more similar DNA, RNA and protein sequences. In 1977, Carl Woese(10) firstly analyzed the phylogenetic relationship based on 16S ribosomal RNA, which became one standard for the research of evolutionary biology based on DNA, RNA and Protein sequences. Later, The multiple sequence alignment (MSA) method

was developed to deal with the current huge data set of DNA, RNA or protein sequences. The distance matrix based on the aligned result is applied to build the phylogenetic tree, which correspond to the evolutionary relationship among the input sequences. There are many multiple sequence alignment computer programs available on internet, such as Clustal(11), Muscle(12) and MAFFT(13). Those programs can receive good alignment result and create accurate phylogenetic tree. However, the computation time will increase rapidly when the number of sequences or the lengths of sequences increase. Hence, we introduce the distribution vectors method in chapter three, which can construct the accurate phylogenetic tree in linear time more faster than the multiple sequence alignment programs.

CHAPTER 2

PREDICTING CODING REGIONS

2.1 Introduction

The DNA sequences are discovered as some kinds of permutations of four nucleotides through biologic experiments, which include Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). Therefore, a DNA sequence can be considered as a character sequence constructed by four letters, A, C, G and T. In order to analyze the functions of DNA sequences in computer aided biological research, the character sequences should be translated into their numerical representations. In the last twenty years, many kind of numerical representations and graphical representations have been provided, such as Gate(14) and Yau(15). In this chapter, We assign the four 4-D unit base vectors to the four nucleotides as follows.

$$\begin{pmatrix}1\\0\\0\\0\end{pmatrix} \to A \qquad \begin{pmatrix}0\\1\\0\\0\end{pmatrix} \to T \qquad \begin{pmatrix}0\\0\\1\\0\end{pmatrix} \to C \qquad \begin{pmatrix}0\\0\\0\\1\end{pmatrix} \to G$$

Then the character sequence of a DNA sequence is translated to a numerical sequence in a 4-dimensional space as

$$x(n) = u_A(n)\begin{pmatrix}1\\0\\0\\0\end{pmatrix} + u_T(n)\begin{pmatrix}0\\1\\0\\0\end{pmatrix} + u_C(n)\begin{pmatrix}0\\0\\1\\0\end{pmatrix} + u_G(n)\begin{pmatrix}0\\0\\0\\1\end{pmatrix}, \tag{2.1}$$

n=0,1,. . . ,N-1

where N is the length of the sequence, and $u_\alpha(n)$ is the indicator sequence

$$u_\alpha(n) = \begin{cases} 1, & \alpha \text{ appears at location n,} \\ 0, & \text{otherwise,} \end{cases} \tag{2.2}$$

where $\alpha \in I = \{A, T, C, G\}$ and n=0,1,. . . ,N-1.

After the numerical sequence is defined, the Fourier transform of the 4-D numerical representation of a DNA sequence at Equation 2.1 can be expressed as follows.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-i\frac{2\pi kn}{N}}, \text{ at frequences } k = 0, 1, \ldots, N-1 \tag{2.3}$$

Similarly,

$$U_\alpha(k) = \sum_{n=0}^{N-1} u_\alpha(n) e^{-i\frac{2\pi kn}{N}}, \text{ at frequences } k = 0, 1, \ldots, N-1, \tag{2.4}$$

where $U_\alpha(k)$ is Fourier transform of the indicator sequences $u_\alpha(k)$. Therefore the power spectrum is

$$P(k) = \bar{X}^T(k)X(k) = \sum_{\alpha \in I} \bar{U}_\alpha(k)U_\alpha(k) \qquad k = 0, 1, \ldots, N-1, \tag{2.5}$$

where $\bar{X}(k)$ is the conjugate of the complex vector $X(k)$ and $\bar{U}_\alpha(k)$ is the conjugate of the complex number $U_\alpha(k)$. After that, we define the total power spectrum as

$$PS = \sum_{k=0}^{N-1} P(k) \tag{2.6}$$

and the average power spectrum as

$$AvgPS = \frac{PS}{N} = \frac{\sum_{k=0}^{N-1} P(k)}{N} \tag{2.7}$$

In 2001, Anastassiou(16) found that the spectrum of a coding DNA sequence usually demonstrates one peak at frequency $k = \frac{N}{3}$ as shown in Figure 3. Later, Vera Afreixo(17) provided a simplified formula to compute the spectrum at frequency $k = \frac{N}{3}$

$$P(\frac{N}{3}) = \sum_{\alpha \in I} \left[(S(0)_\alpha - \frac{S(1)_\alpha + S(2)_\alpha}{2})^2 + \frac{3}{4}(S(1)_\alpha - S(2)_\alpha)^2 \right], \tag{2.8}$$

where $I = \{A, T, C, G\}$ and

$$S(m)_\alpha = \sum_{k=0}^{\frac{N}{3}-1} u_\alpha(3k+m), \qquad m = 0, 1, 2 \tag{2.9}$$

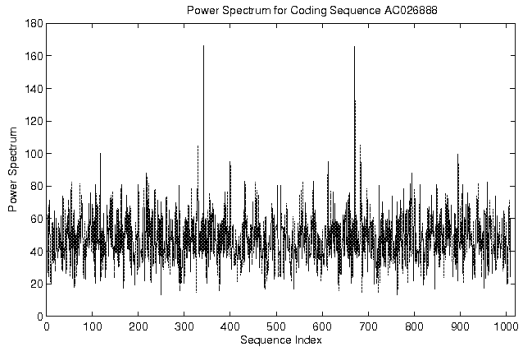

Figure 3. Power Spectrum of One Coding Sequence

Then we can define the parameter $TICOR$ (Threshold to Identify Coding Regions) as the ratio of $P(\frac{N}{3})$ to the average power spectrum $avgPS$ as follows

$$TICOR = \frac{P(\frac{N}{3})}{avgPS} \tag{2.10}$$

$$= \frac{\sum_{\alpha \in I} \left[(S(0)_\alpha - \frac{S(1)_\alpha + S(2)_\alpha}{2})^2 + \frac{3}{4}(S(1)_\alpha - S(2)_\alpha)^2 \right]}{\frac{\sum_{k=0}^{N-1} P(k)^2}{N}} \tag{2.11}$$

2.2 Materials and Methods

Theorem 1: The sum of power spectrum is N^2

$$PS = \sum_{k=0}^{N-1} P(k) = \sum_{k=0}^{N-1} \sum_{\alpha \in I} \bar{U}_\alpha(k) U_\alpha(k) = N^2, \tag{2.12}$$

where $\alpha \in I = \{A, T, C, G\}$.

Proof. At first, we know

$$\bar{U}_\alpha(k)U_\alpha(k) = \left(\overline{\sum_{l=0}^{N-1} u_\alpha(l)e^{-i\frac{2\pi}{N}lk}}\right)\left(\sum_{j=0}^{N-1} u_\alpha(j)e^{-i\frac{2\pi}{N}jk}\right)$$
$$= \sum_{l\in S_\alpha} e^{i\frac{2\pi}{N}lk}\sum_{j\in S_\alpha} e^{-i\frac{2\pi}{N}jk} = |S_\alpha| + \sum_{l,j\in S_\alpha\ l\neq j} e^{-i\frac{2\pi}{N}(j-l)k}, \tag{2.13}$$

where $S_\alpha = \{n|u_\alpha(n) = 1, \quad n = 0, 1, ..., N-1\}, \quad \alpha \in I = \{A, T, C, G\}$. It is clear that

$$S_A \bigcup S_T \bigcup S_G \bigcup S_C = \{0, 1, \ldots, N-1\} \tag{2.14}$$

and

$$|S_A| + |S_T| + |S_C| + |S_G| = N \tag{2.15}$$

Therefore,

$$\sum_{k=0}^{N-1} \bar{U}_\alpha(k)U_\alpha(k) = N|S_\alpha| + \sum_{k=0}^{N-1}\sum_{l,j\in S_\alpha\ l\neq j} e^{-i\frac{2\pi}{N}(j-l)k}$$
$$= N|S_\alpha| + \sum_{l,j\in S_\alpha\ l\neq j}\sum_{k=0}^{N-1} e^{-i\frac{2\pi}{N}(j-l)k} \tag{2.16}$$

Here we can obtain the following result by using the geometric series

$$\sum_{k=0}^{N-1} e^{-i\frac{2\pi}{N}(l-j)k} = \frac{1-e^{-i\frac{2\pi}{N}(l-j)N}}{1-e^{-i\frac{2\pi}{N}(l-j)}} = \frac{1-1}{1-e^{-i\frac{2\pi}{N}(l-j)}} = 0, \quad i \neq j \tag{2.17}$$

Therefore,

$$\sum_{k=0}^{N-1} \bar{U}_\alpha(k) U_\alpha(k) = N|S_\alpha| \tag{2.18}$$

Finally,

$$PS = \sum_{k=0}^{N-1} P(k) = \sum_{k=0}^{N-1} \sum_{\alpha \in I} \bar{U}_\alpha(k) U_\alpha(k) = \sum_{\alpha \in I} N|S_\alpha| = N^2 \tag{2.19}$$

□

Obviously,the average power spectrum $avgPS = N$. Therefore,

$$TICOR = \frac{\sum_{\alpha \in I} \left[(S(0)_\alpha - \frac{S(1)_\alpha + S(2)_\alpha}{2})^2 + \frac{3}{4}(S(1)_\alpha - S(2)_\alpha)^2 \right]}{N} \tag{2.20}$$

It is clear that $P(\frac{N}{3}) \geq 0$, and $P(\frac{N}{3}) = 0$ if and only if $S(0)_\alpha = S(1)_\alpha = S(2)_\alpha$. So $P(\frac{N}{3})$ and $TICOR$ have the minimum value zero when $S(0)_\alpha = S(1)_\alpha = S(2)_\alpha, \quad \alpha \in I, I = A, C, G, T$. On the other hand, Equation 2.8 can be written as follows,

$$\begin{aligned}
P(\frac{N}{3}) &= \sum_{\alpha \in I} \left[(S(0)_\alpha - \frac{S(1)_\alpha + S(2)_\alpha}{2})^2 + \frac{3}{4}(S(1)_\alpha - S(2)_\alpha)^2 \right] \\
&= \sum_{\alpha \in I} \left[S(0)_\alpha^2 + S(1)_\alpha^2 + S(2)_\alpha^2 - S(0)_\alpha S(1)_\alpha - S(0)_\alpha S(2)_\alpha - S(1)_\alpha S(2)_\alpha \right] \\
&= \sum_{l=0}^{2} (S(l)_A^2 + S(l)_T^2 + S(l)_C^2 + S(l)_G^2) - \sum_{\alpha \in I} \sum_{\substack{j=0 \\ j \neq k}}^{2} \sum_{k=0}^{2} S(j)_\alpha S(k)_\alpha \\
&\leq \sum_{l=0}^{2} (S(l)_A^2 + S(l)_T^2 + S(l)_C^2 + S(l)_G^2) \\
&= \sum_{l=0}^{2} (S(l)_A + S(l)_T + S(l)_C + S(l)_G)^2 - \sum_{l=0}^{2} \sum_{\substack{\alpha, \beta \in I \\ \alpha \neq \beta}} S(l)_\alpha S(l)_\beta \\
&\leq \sum_{l=0}^{2} (S(l)_A + S(l)_T + S(l)_C + S(l)_G)^2 \\
&= 3(\frac{N}{3})^2 = \frac{N^2}{3}
\end{aligned} \tag{2.21}$$

Therefore, $P(\frac{N}{3})$ has the maximum value $\frac{N^2}{3}$ and $TICOR$ has maximum value $\frac{N}{3}$ when

$$S(j)_\alpha S(k)_\alpha = 0, \quad j, k = 0, 1, 2, j \neq k \quad \alpha \in I \tag{2.22}$$

and

$$S(l)_\alpha S(l)_\beta = 0, \quad \alpha, \beta \in I, \alpha \neq \beta, l = 0, 1, 2 \tag{2.23}$$

It means a DNA sequence has the maximum value of the function $P(\frac{N}{3})$ when the sequence is constructed by a same codon repeated, such as ACT ACT ACT...ACT.

When we assume that each nucleotide in one DNA sequence could be A,C,G or T with the probability $\frac{1}{4}$ randomly and independently, all the expectations of $S(k)_\alpha$ are $\frac{N}{12}$. Therefore, we can get the expectation of $P(\frac{N}{3})$ as follows.

$$\begin{aligned}
&E[P(\frac{N}{3})] \\
&= E[\sum_{\alpha \in I}(S(0)_\alpha^2 + S(1)_\alpha^2 + S(2)_\alpha^2 - S(0)_\alpha S(1)_\alpha - S(0)_\alpha S(2)_\alpha - S(1)_\alpha S(2)_\alpha)] \\
&= 12(E[S(0)_\alpha^2] - E[S(0)_\alpha]^2) = 12Var[S(0)_\alpha] \\
&= 12\frac{N}{3}\frac{1}{4}(1 - \frac{1}{4}) = \frac{3N}{4}
\end{aligned} \tag{2.24}$$

Then we can get the expectation of $TICOR$

$$E[TICOR] = \frac{3}{4} = 0.75 \tag{2.25}$$

In order to predict the coding regions in the DNA sequence, we compute the average of $TICOR$ for all $N - w + 1$ subsequences by sliding a window with a fixed size w on the DNA sequence from the first nucleotide to the end, which we call $avgTICOR$. The fixed size w should be a

multiple of three to compute the $TICOR$. In this thesis, we choose $w = 102$ to optimize the accuracy of predicting the coding regions.

$$avgTICOR = \frac{\sum_{k=1}^{N-w+1} TICOR(seq(k : k + w - 1))}{N - w + 1}, \tag{2.26}$$

where $TICOR(seq(k : k + w - 1))$ means the value of $TICOR$ of the subsequence that starts from the position k and stop at the position $k + w - 1$

Therefore,

$$E[avgTICOR] = \frac{\sum_{k=1}^{N-w+1} E[TICOR(seq(k : k + w - 1))]}{N - w + 1} = 0.75 \tag{2.27}$$

Moreover, we define $\overline{TICOR(i)}$ and $V(i)$ to represent the property of the nucleotide at the position i.

$$\overline{TICOR(i)} = \begin{cases} \dfrac{\sum_{k=i-w+1}^{i} TICOR(seq(k : k + w - 1))}{w} \\ \qquad i = w, w + 1, ..., N - w, N - w + 1 \\ \overline{TICOR(w)} \\ \qquad i = 1, 2, ..., w - 2, w \\ \overline{TICOR(N - w + 1)} \\ \qquad i = N - w + 2, N - w + 3, ..., N - 1, N \end{cases} \tag{2.28}$$

$$V(i) = \begin{cases} \frac{(\overline{TICOR(i-5)} - \overline{TICOR(i+5)})^2}{\overline{TICOR(i)}^2} \\ \qquad i = w-4, w-3, ..., N-w+4, N-w+5 \\ 0 \qquad i = 1, 2, ...w-5 \text{ or } N-w+6, ..., N-1, N \end{cases} \tag{2.29}$$

2.3 Application

At first, we calculate the value of $avgTICOR$ for 500 human coding sequences and 500 human non-coding sequences randomly chosen from NCBI database. Figure 4 shows that there are huge differences between the value of $avgTICOR$ of coding sequences and non-coding sequences. The value of $avgTICOR$ for non-coding sequences are close to the mathematical expectation of $avgTICOR$ for random sequences, whereas most coding sequences hold much higher value of $avgTICOR$.

Secondly, we compute the value of $avgTICOR$ and the proportion of coding regions to the length of the whole DNA sequence for 200 human DNA sequences. These 200 pairs of data ($avgTICOR$, proportion) are plotted in Figure 5. Then the linear equation of the proportion depending on $avgTICOR$ is estimated by using the least squares method.

$$proportion = 32.1827 \times avgTICOR - 7.7358 \tag{2.30}$$

Finally, we describe the strategy to predict the coding regions as follows, which we call TICORSCAN.

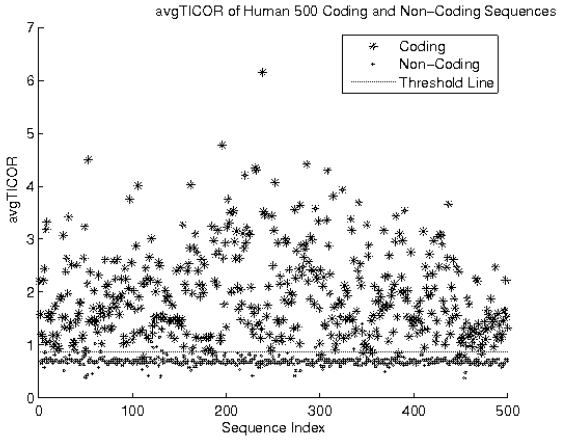

Figure 4. avgTICOR of 500 Human Coding Sequences and Non-Coding Sequences

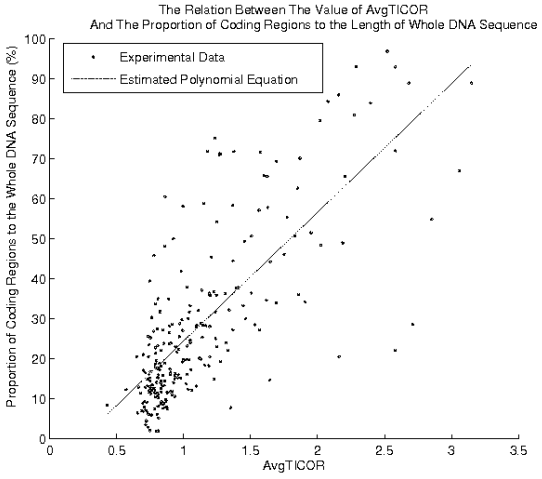

Figure 5. The Relation Between avgTICOR and The Proportion of Coding Regions to the Length of Whole DNA Sequences

1. Calculate $avgTICOR$, $\overline{TICOR(i)}$ and $V(i)$ for the DNA sequence.

2. Estimate the *proportion* of coding regions to the whole DNA sequence by using Equation 2.30. Then determine the threshold line, which is horizontal line. The regions where the $\overline{TICOR}$ curve above this horizontal line are the potential coding regions. The horizontal line is determined by the proportion of the coding regions.

3. Detect the points of intersection between the curve of $\overline{TICOR}$ and the threshold line.

4. Consider the points of intersection as above. For the point with a neighborhood where the $\overline{TICOR}$ curve is increasing, this point is called the start position of the coding region. Similarly, for the point with a neighborhood where the $\overline{TICOR}$ curve is decreasing, this point is called the end position of the coding region.

5. Label the first position of the DNA sequence as the first start position of coding regions if $\overline{TICOR(1)}$ is above the threshold line, the last position of the DNA sequence as the last end position of coding regions if $\overline{TICOR(N)}$ is above the threshold line.

6. Relocate the start and end position in their neighborhood. We set the position which has the maximum value of $V(i)$ as the start or end position. The neighborhood is defined as (position-25,position+25) in this application. Moreover, decide the first start position and the last end position by considering the biology information: The start codon is ATG and the stop codon is TGA,TAA or TAG.

7. Remove the pair of start and end position if the distance between them is too short, which means the distance is less than 50 in this application. This step corrects the prediction of short non-coding region over the threshold line.

8. Remove the end position and the next start position if the distance between them is too short, which means the distance is less than 50 in this application. This step corrects the prediction of short coding region below the threshold line.

9. Pair the start positions and the end positions of coding regions to finish predicting the coding regions.

We apply our TICORSCAN method into the ROSETTA dataset which includes 117 orthologous human and mouse DNA sequences. Figure 6, Figure 7, Figure 8 and Figure 9 display the process to predict the coding regions from the DNA sequence M11160. More examples are illustrated in Figure 10, Figure 11, Figure 12 and Figure 13. Especially, Figure 12 shows that there is a short non-coding region above the threshold line, which is successfully predicted as the non-coding region by following the rule 7. On the other hand, there is a short coding region below the threshold line in Figure 13. This region is considered as the coding region by applying the rule 8. However, these two rules mistake the prediction of short non-coding regions and coding regions sometimes. Over all, these two rules are important to increase the prediction accuracy since the short non-coding and coding regions only constitute a small percentage of the total nucleotides.

In order to compare the prediction accuracy, we predict the coding regions in the same dataset by using GENSCAN, TBLASTX and other programs. To evaluate the accuracy, we

count TP (true positives) as the number of nucleotides in the coding regions that are predicted as in the coding regions, FP (false positives) as the number of nucleotides in the non-coding regions which are predicted as in the coding regions, FN (false negatives) as the number of nucleotides in the coding regions that are predicted as in the non-coding regions and TN (true negatives) as the number of nucleotides in the non-coding regions which are predicted as in the non-coding regions. Then we define sensitivity (Sn), specificity (Sp) and the approximate correlation (AC) that summarizes the overall nucleotide sensitivity and specificity by one number. We compute the overall prediction result, which includes Sn, Sp and AC, for our TICORSCAN method and other programs. The accuracy in Table I shows that our TICORSCAN method performs very well on the coding regions prediction. Moreover, we divide the ROSETTA dataset into five subgroups by the range of the length of DNA sequences. The prediction record of the five subgroups listed in Table II shows that our TICORSCAN method has a better prediction result on the long coding regions.

$$Sn = \frac{TP}{TP + FN} \tag{2.31}$$

$$Sp = \frac{TP}{TP + FP} \tag{2.32}$$

$$AC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right) - 1 \tag{2.33}$$

TABLE I

PREDICTION ACCURACY ON THE ROSETTA DATASET

Program	Sn	Sp	AC
GENSCAN	97.5	90.8	92.9
TBLASTX default	94.0	80.3	88.1
TWINSCAN	98.4	88.9	92.3
SGP-1	94.0	96.0	94.0
TICORSCAN	96.7	92.6	94.1

TABLE II

STATISTICS OF THE SUBSETS GROUPED BY THE LENGTH OF ACTUAL CODING REGIONS

Length Range	Number of Coding Regions		Number of the Nucleotides	Percentage
	Actual	Predicted		
< 50	82	35	2488	1.1%
≥ 50 & < 100	209	134	15564	6.89%
≥ 100 & < 150	304	282	37974	16.76%
≥ 150 & < 250	249	243	46251	20.47%
≥ 250	189	188	123667	54.73%
Total	1033	847	225944	100%

2.4 Conclusion

This chapter introduces a parameter $TICOR$ derived from the spectrum at frequency $k = \frac{N}{3}$. The $TICOR$ presents the difference between coding regions and non-coding regions and shows that the non-coding regions has the similar mathematical properties with the random sequences. Moreover, we provide a linear equation to estimate the proportion of coding regions

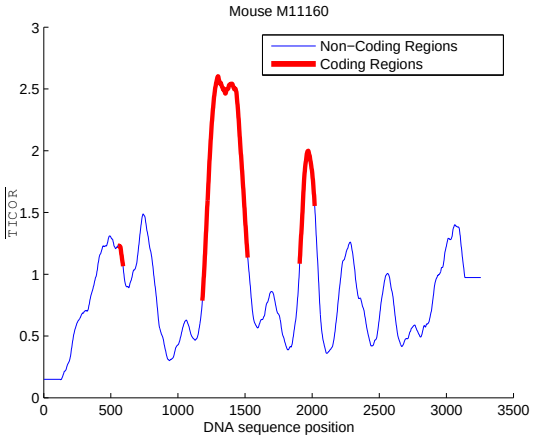

Figure 6. Drawing the $\overline{TICOR}$ Curve

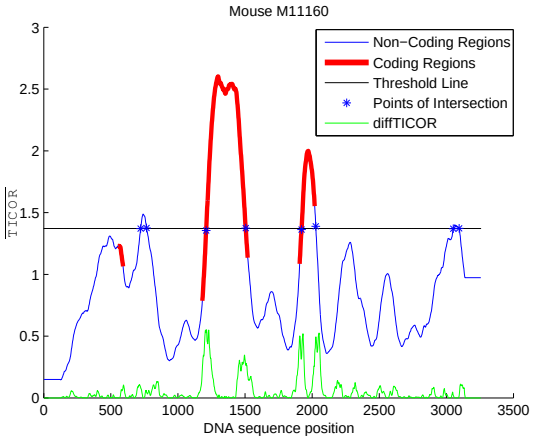

Figure 7. Finding the points of intersection

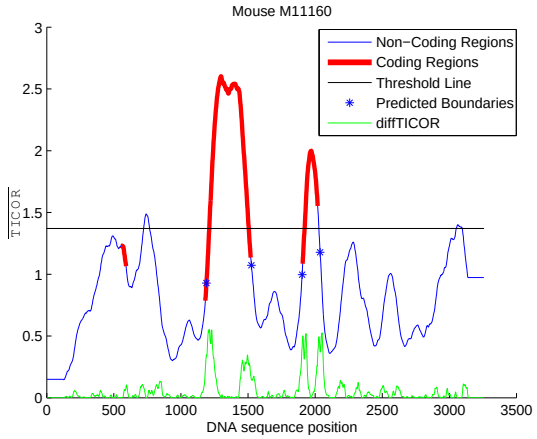

Figure 8. Predicting the start and end positions

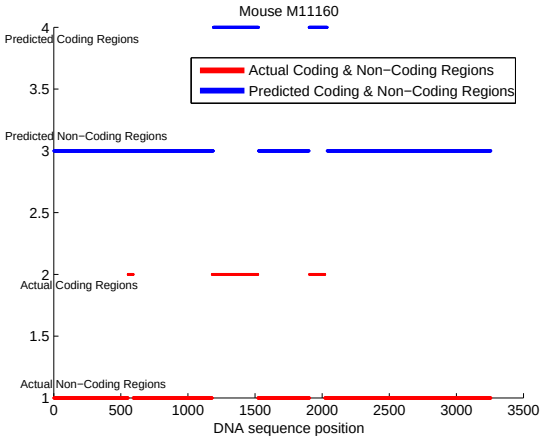

Figure 9. Predicted Result of Mouse M11160

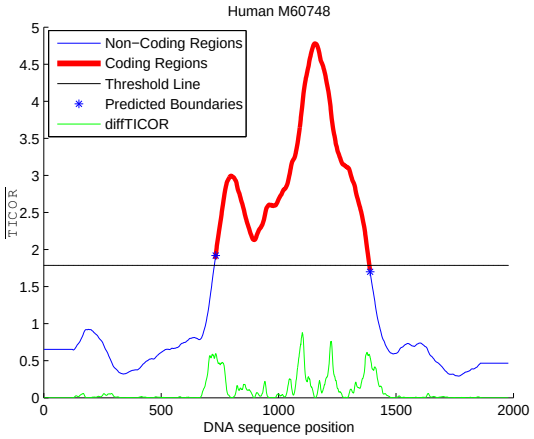

Figure 10. Single Coding Region

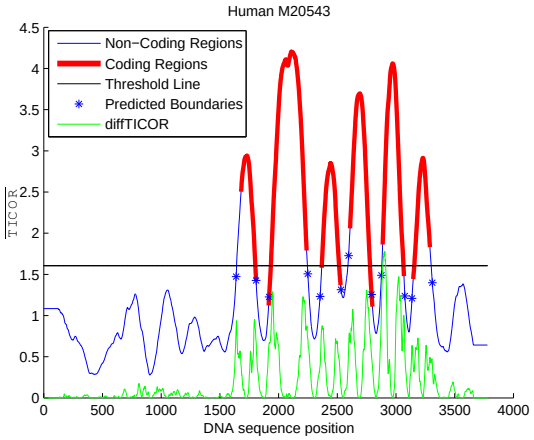

Figure 11. Multiple Coding Regions

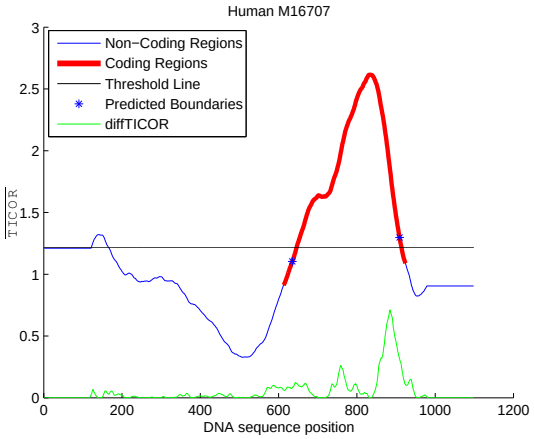

Figure 12. Short Non-Coding Region over Threshold

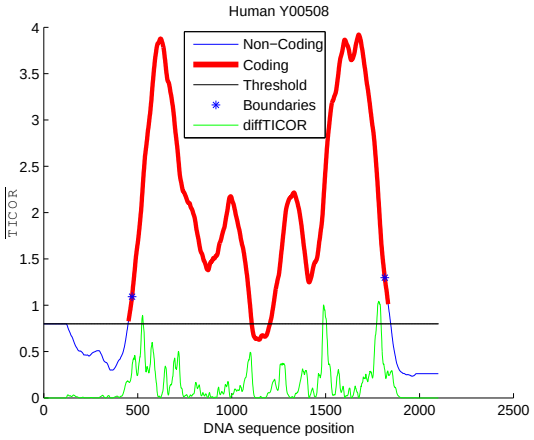

Figure 13. Short Coding Region below Threshold

to the length of the whole DNA sequence based on its $avgTICOR$ value. Finally, we develop a new method to predict the coding regions in the DNA sequences, which we call TICORSCAN. Our method is a faster method with linear computation time compared to other popular methods based on statistics model, machine learning or alignment method. Moreover, unlike these complicated methods, our TICORSCAN method is straightforward to understand and easy to implement. We test our method on the ROSETTA dataset. The experiment obtains the high accuracy to predict the coding regions. Especially, our method works better for the long coding regions, which successfully predict almost all the coding regions whose length is longer than 150. However, the TICORSCAN method is not good for the short coding regions. First, many short coding regions do not have obvious peaks in their $\overline{TICOR}$ curves. Secondly, in order to keep the overall high accuracy, we consider the short regions which are less than 50 as the non-coding regions even if they have peaks. Fortunately, these missing short coding regions affect the prediction accuracy little because the proportion of the nucleotides of missing short coding regions to the all coding regions is tiny. Over all, our TICORSCAN method is an efficient method to predict the coding regions with high accuracy.

CHAPTER 3

CLUSTERING HOMOLOGOUS DNA SEQUENCES

3.1 Methods

In the beginning, we define the indicator sequence $u_\alpha(n)$ of the DNA sequence.

$$u_\alpha(n) = \begin{cases} 1, & \text{if } \alpha \text{ appears at location n of the DNA sequence,} \\ 0, & \text{otherwise,} \end{cases} \tag{3.1}$$

$\alpha \in I = \{A, T, C, G\}$, n=0,1,. . . ,N-1 and N is the length of the DNA sequence.

To construct the distribution vectors, we fix k, which is a preset integer much less than N. Then we define q as the quotient and r as the remainder in Equation 3.2 when dividing N by k.

$$q = \lfloor \frac{N}{k} \rfloor, \quad r = N - k \times q \tag{3.2}$$

It is clear that $0 \leq r < k$. Therefore, we divide the DNA sequences into k segments with almost equal lengths: The first r segments possess q+1 nucleotides and the remaining k-r segments hold q nucleotides. Equation 3.3 explains the partition clearly.

$$N = k \times q + r = r(q+1) + (k-r)q \tag{3.3}$$

Then we define $Q_\alpha(m,k)$ as the number of the nucleotides α in the m^{th} segment of the DNA sequence in Equation 3.4.

$$Q_\alpha(m,k) = \begin{cases} \sum_{i=m(q+1)}^{m(q+1)+q} u_\alpha(i), & m = 0,1,2...r-1 \\ \\ \sum_{i=m\times q+r}^{(m+1)q+r-1} u_\alpha(i), & m = r, r+1...k-1 \end{cases} \tag{3.4}$$

For each k, we define the $DV_\alpha(k)$ in terms of $Q_\alpha(m,k)$ to describe the variability between any two of $Q_\alpha(m,k)$ for the particular nucleotide α in one DNA sequence.

$$DV_\alpha(k) = \frac{8}{3N(k-1)} \left(\sum_{\substack{i=0 \\ i\neq j}}^{k-1} \sum_{j=0}^{k-1} (Q_\alpha(i,k) - Q_\alpha(j,k))^2 \right) \tag{3.5}$$

The intention for choosing the coefficient $\frac{8}{3N(k-1)}$ is to simplify the expectation to be a constant. The explanation will be given later.

For each $k \in K = \{3,4,5,7,11,13,17,19,23,29,31,37,41,43,47\}$, we compute the $DV_A(k)$, $DV_C(k)$, $DV_G(k)$ and $DV_T(k)$ and put these together to obtain the sixty dimensional distribution vector $\overline{DV}$. It is clear there is no common factor except 1 among the numbers in the set K, which makes the elements in the distribution vector more independent. The selection of the size of the set K is crucial. The distribution vectors can map the sequences more precisely when the size of K is large, which consists of only prime numbers except 4. On the other hand for the short sequences, each segment is too short to provide the information if the k is too

large. In addition, the larger the size of the set K, the longer the computation time. All of the above reasons should be considered in the selection of the set K.

$$\begin{aligned}\overline{DV} = \{&DV_A(3), DV_C(3), DV_G(3), DV_T(3),\\ &DV_A(4), DV_C(4), DV_G(4), DV_T(4),\\ &\ldots\\ &DV_A(47), DV_C(47), DV_G(47), DV_T(47)\}\end{aligned} \tag{3.6}$$

3.2 Statistical Properties

In order to study the expectation and standard deviation of the distribution vectors, we need to consider the DNA sequence as a random sequence, which means every position in the DNA sequence can be A, C, G or T with the same probability $\frac{1}{4}$ independently.

First, we compute the expectation of $Q_\alpha(m,k)$, $Q_\alpha^2(m,k)$, $Q_\alpha^3(m,k)$ and $Q_\alpha^4(m,k)$ and the variance of $Q_\alpha(m,k)$ respectively.

$$E[Q_\alpha(m,k)] = \sum_{i=0}^{n}(i\frac{n!}{i!(n-i)!}p^i(1-p)^{n-i}) = \frac{N}{4k} \tag{3.7}$$

$$\begin{aligned}E[Q_\alpha^2(m,k)] &= \sum_{i=0}^{n}(i^2\frac{n!}{k!(n-i)!}p^i(1-p)^{n-i})\\ &= pn(pn-p+1) = \frac{N}{4k}(\frac{N}{4k}+\frac{3}{4})\end{aligned} \tag{3.8}$$

$$
\begin{aligned}
Var[Q_\alpha(m,k)] &= E[Q_\alpha^2(m,k)] - E^2[Q_\alpha(m,k)] \\
&= \frac{N}{4k}(\frac{N}{4k} + \frac{3}{4}) - (\frac{N}{4k})^2 = \frac{3N}{16K}
\end{aligned} \tag{3.9}
$$

$$
\begin{aligned}
E[Q_\alpha^3(m,k)] &= \sum_{i=0}^{n}(i^3 \frac{n!}{k!(n-i)!} p^i (1-p)^{n-i}) \\
&= pn(p^2n^2 - 3p^2n + 2p^2 + 3pn - 3p + 1) \\
&= \frac{N}{4k}(\frac{N^2}{16k^2} + \frac{9N}{16k} + \frac{3}{8})
\end{aligned} \tag{3.10}
$$

$$
\begin{aligned}
E[Q_\alpha^4(m,k)] &= \sum_{i=0}^{n}(i^4 \frac{n!}{k!(n-i)!} p^i (1-p)^{n-i}) \\
&= pn(p^3n^3 + 6p^2n^2 - 6p^3n^2 - 18p^2n \\
&\quad + 11p^3n + 7pn - 6p^3 + 12p^2 - 7p + 1) \\
&= \frac{N}{4k}(\frac{N^3}{64k^3} + \frac{9N^2}{32k^2} + \frac{51N}{64k} - \frac{3}{32}) \\
&\text{where } n = \frac{N}{k} \text{ and } p = \frac{1}{4}
\end{aligned} \tag{3.11}
$$

Now we can compute the expectation and standard deviation of $DV_\alpha(k)$.

$$
\begin{aligned}
E[DV_\alpha(k)] &= E\Big[\frac{8}{3N(k-1)}\Big(\sum_{\substack{i=0\\ i\neq j}}^{k-1}\sum_{j=0}^{k-1}(Q_\alpha(i,k)-Q_\alpha(j,k))^2\Big)\Big] \\
&= E\Big[\frac{16}{3N}\Big(\sum_{m=0}^{k-1}Q_\alpha^2(m,k)-\frac{1}{k-1}\sum_{i=0,i\neq j}^{k-1}\sum_{j=0}^{k-1}Q_\alpha(i,k)Q_\alpha(j,k)\Big)\Big] \\
&= \frac{16}{3N}\Big(\sum_{m=0}^{k-1}E[Q_\alpha^2(m,k)]-\frac{1}{k-1}\sum_{i=0,i\neq j}^{k-1}\sum_{j=0}^{k-1}E[Q_\alpha(i,k)]E[Q_\alpha(j,k)]\Big) \\
&= \frac{16k}{3N}(E[Q_\alpha^2(m,k)]-E[Q_\alpha(i,k)]E[Q_\alpha(j,k)]) \\
&= \frac{16k}{3N}Var[Q_\alpha(m,k)] \\
&= \frac{16k}{3N}\frac{3N}{16K} \\
&= 1
\end{aligned}
\tag{3.12}
$$

And

$$E[DV_\alpha^2(k)] = E\Big[\Big(\frac{8}{3N(k-1)}\Big(\sum_{\substack{i=0\\i\neq j}}^{k-1}\sum_{j=0}^{k-1}(Q_\alpha(i,k)-Q_\alpha(j,k))^2\Big)\Big)^2\Big]$$

$$= E\Big[\frac{256}{9N^2}\Big(\sum_{m=0}^{k-1}Q_\alpha^2(m,k) - \frac{1}{k-1}\sum_{\substack{i=0\\i\neq j}}^{k-1}\sum_{j=0}^{k-1}Q_\alpha(i,k)Q_\alpha(j,k)\Big)^2\Big]$$

$$= \frac{256}{9N^2(k-1)^2}E\Big[\Big((k-1)\sum_{m=0}^{k-1}Q_\alpha^2(m,k) - \sum_{\substack{i=0\\i\neq j}}^{k-1}\sum_{j=0}^{k-1}Q_\alpha(i,k)Q_\alpha(j,k)\Big)^2\Big]$$

$$= \frac{256}{9N^2(k-1)^2}E\Big[(k-1)^2(\sum_{m=0}^{k-1}Q_\alpha^2(m,k))^2$$
$$- 2(k-1)(\sum_{m=0}^{k-1}Q_\alpha^2(m,k))(\sum_{\substack{i=0\\i\neq j}}^{k-1}\sum_{j=0}^{k-1}Q_\alpha(i,k)Q_\alpha(j,k)) + (\sum_{\substack{i=0\\i\neq j}}^{k-1}\sum_{j=0}^{k-1}Q_\alpha(i,k)Q_\alpha(j,k))^2\Big]$$

$$= \frac{256}{9N^2(k-1)^2}E\Big[(k-1)^2\big(\sum_{m=0}^{k-1}Q_\alpha^4(m,k) + \sum_{\substack{i=0\\i\neq j}}^{k-1}\sum_{j=0}^{k-1}Q_\alpha^2(i,k)Q_\alpha^2(j,k)\big)$$
$$- 4(k-1)\sum_{\substack{i=0\\i\neq j}}^{k-1}\sum_{j=0}^{k-1}Q_\alpha^3(i,k)Q_\alpha(j,k) - 2(k-1)\sum_{\substack{i=0\\i\neq j,l}}^{k-1}\sum_{\substack{j=0\\j\neq l}}^{k-1}\sum_{l=0}^{k-1}Q_\alpha^2(i,k)Q_\alpha(j,k)Q_\alpha(l,k)$$
$$+ 2\sum_{\substack{i=0\\i\neq j}}^{k-1}\sum_{j=0}^{k-1}Q_\alpha^2(i,k)Q_\alpha^2(j,k) + 4\sum_{\substack{i=0\\i\neq j,l}}^{k-1}\sum_{\substack{j=0\\j\neq l}}^{k-1}\sum_{l=0}^{k-1}Q_\alpha^2(i,k)Q_\alpha(j,k)Q_\alpha(l,k)$$
$$+ \sum_{\substack{i=0\\i\neq j,l,m}}^{k-1}\sum_{\substack{j=0\\j\neq l,m}}^{k-1}\sum_{\substack{l=0\\l\neq m}}^{k-1}\sum_{m=0}^{k-1}Q_\alpha(i,k)Q_\alpha(j,k)Q_\alpha(l,k)Q_\alpha(m,k)\Big]$$

$$= \frac{256}{9N^2(k-1)^2}E\Big[(k-1)^2\sum_{m=0}^{k-1}Q_\alpha^4(m,k) - 4(k-1)\sum_{\substack{i=0\\i\neq j}}^{k-1}\sum_{j=0}^{k-1}Q_\alpha^3(i,k)Q_\alpha(j,k)$$
$$+ ((k-1)^2+2)\sum_{\substack{i=0\\i\neq j}}^{k-1}\sum_{j=0}^{k-1}Q_\alpha^2(i,k)Q_\alpha^2(j,k) - 2(k-3)\sum_{\substack{i=0\\i\neq j,l}}^{k-1}\sum_{\substack{j=0\\j\neq l}}^{k-1}\sum_{l=0}^{k-1}Q_\alpha^2(i,k)Q_\alpha(j,k)Q_\alpha(l,k)$$
$$+ \sum_{\substack{i=0\\i\neq j,l,m}}^{k-1}\sum_{\substack{j=0\\j\neq l,m}}^{k-1}\sum_{\substack{l=0\\l\neq m}}^{k-1}\sum_{m=0}^{k-1}Q_\alpha(i,k)Q_\alpha(j,k)Q_\alpha(l,k)Q_\alpha(m,k)\Big]$$

$$= \frac{256}{9N^2(k-1)^2}\Big(k(k-1)^2E[Q_\alpha^4(m,k)] - 4k(k-1)^2E[Q_\alpha^3(m,k)]E[Q_\alpha(m,k)]$$
$$+ k(k-1)((k-1)^2+2)E^2[Q_\alpha^2(m,k)] - 2k(k-1)(k-2)(k-3)E[Q_\alpha^2(m,k)]E^2[Q_\alpha(m,k)]$$
$$+ k(k-1)(k-2)(k-3)E^4[Q_\alpha(m,k)]\Big) \tag{3.13}$$

Combining the results from Equation 3.7, Equation 3.8, Equation 3.10 and Equation 3.11 into Equation 3.13.

$$\begin{aligned} E[DV_\alpha^2(k)] = {} & \frac{256}{9N^2(k-1)^2}(k(k-1)^2\frac{N}{4k}(\frac{N^3}{64k^3} + \frac{9N^2}{32k^2} + \frac{51N}{64k} - \frac{3}{32}) \\ & - 4k(k-1)^2\frac{N}{4k}(\frac{N^2}{16k^2} + \frac{9N}{16k} + \frac{3}{8})\frac{N}{4k} \\ & + k(k-1)((k-1)^2+2)(\frac{N}{4k}(\frac{N}{4k} + \frac{3}{4}))^2 \\ & - 2k(k-1)(k-2)(k-3)\frac{N}{4k}(\frac{N}{4k} + \frac{3}{4})(\frac{N}{4k})^2 \\ & + k(k-1)(k-2)(k-3)E(\frac{N}{4k})^4) \\ = {} & \frac{k+1}{k-1} - \frac{2}{3N} \end{aligned} \tag{3.14}$$

Therefore,

$$Var[DV_\alpha(k)] = E[DV_\alpha^2(k)] - E^2[DV_\alpha(k)] \tag{3.15}$$
$$= \frac{k+1}{k-1} - \frac{2}{3N} - 1 = \frac{2}{k-1} - \frac{2}{3N}$$
$$\approx \frac{2}{k-1}, \qquad \text{when N is large.} \tag{3.16}$$

Then,

$$std[DV_\alpha(k)] = \sqrt{Var[DV_\alpha(k)]} \approx \sqrt{\frac{2}{k-1}} \tag{3.17}$$

After deriving the equations of the expectation and standard deviation of the distribution vectors, we examine these properties with two large datasets. One is 5000 Human DNA sequences from the NCBI database, which are divided into five groups by the respective lengths of DNA sequences. The detail of grouping is provided in Table III. Another dataset is 5000 random DNA sequences which are divided into five groups also. Each group consists of 1000 random sequences each with a fixed length. The lengths of these groups are 200, 400, 800, 1500 and 3000 corresponding to the Group I, II, III, IV and V respectively. We compare the means and standard deviations for the ten groups with the theoretical expectation and standard deviation as in Figure 14, Figure 16, Figure 15 and Figure 17. The selection of the set K plays an important role in the values of $DV_\alpha(k)$. The variability of the distribution vectors is small when k is large. Moreover, the mean and standard deviation within the five human DNA sequences groups also converge to the theoretical expectation and standard deviation when we increase the dimension, even though the convergence is not as good as those of the random sequences.

3.3 Application

We apply our distribution vector method to two datasets. One is 80 Mitochondrial complete genomes of placental mammals from NCBI database, another is the gene Haemagglutinin (HA) of 60 H1N1 viruses from Influenza Virus Sequence Database. At first we calculate the distribution vectors of these sequences and the distances between any two of these distribution

TABLE III

THE GROUPING OF 5000 HUMAN DNA SEQUENCES

	Number	Range of length
Group I	996	< 384
Group II	1001	≥ 384 and < 651
Group III	999	≥ 651 and < 1053
Group IV	1500	≥ 1053 and < 2265
Group V	504	≥ 2265

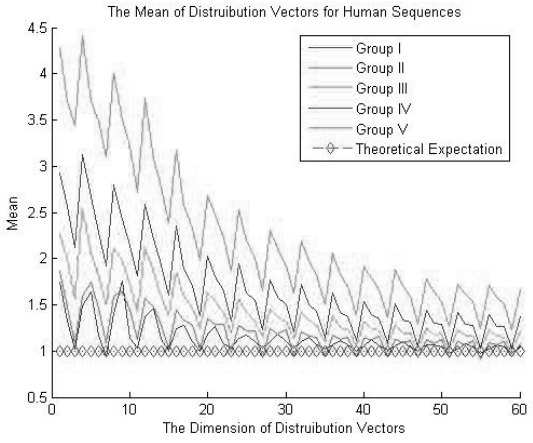

Figure 14. The Mean of Distribution Vectors for Human Sequences

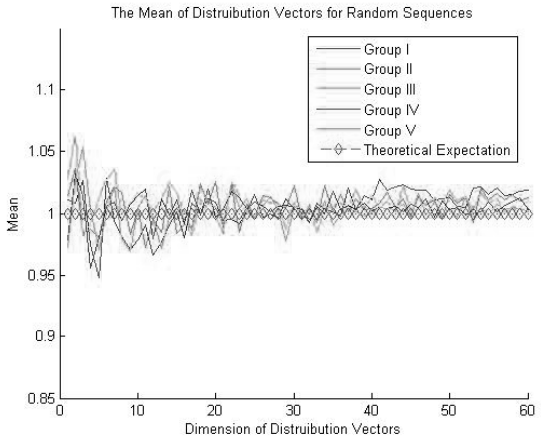

Figure 15. The Mean of Distribution Vectors for Random Sequences

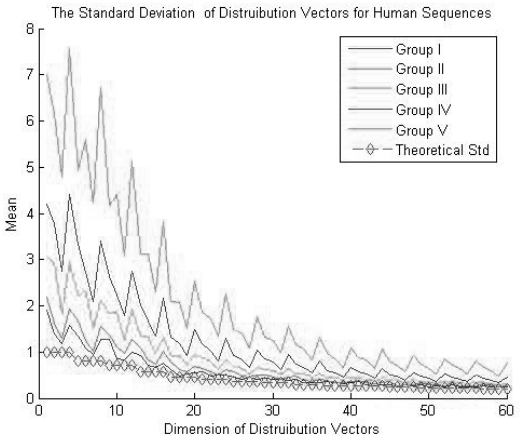

Figure 16. The Standard Deviation of Distribution Vectors for Human Sequences

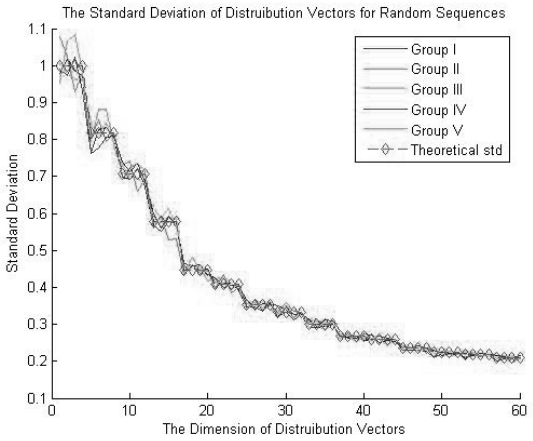

Figure 17. The Standard Deviation of Distribution Vectors for Random Sequences

vectors for each dataset. The phylogenetic trees are built based on the distance matrix by using the function hclust from the R program(18), where the average linkage method is used in the clustering. The two trees are plotted in Figure 20 and Figure 24. Moreover, we apply the multiple alignment on the same two datasets with ClustalW2, MAFFT and Muscle and do the clustering with the average linkage method also. The results are provided in the Supplement. For the dataset of 60 H1N1 (HA) viruses, our distribution vector method classifies these viruses into four groups correctly. The four groups include the avian older than 2009, European swine older than 2009, American swine older than 2009 and the new 2009 viruses from human, swine and avian. The result shows the 2009 human H1N1 viruses have closer relationship with old American swine than old avian and European swine. ClustalW2 and Muscle also classify the 60 H1N1 viruses into the four groups except that the virus swine/wisconsin/1961 is not classified well. Unfortunately, MAFFT is unable to put the old avian viruses into one group. On the

other hand, all the four methods classify most of the 80 animals correctly by the respective orders they belong. Our distribution vector method divides the animals in the order of Carnivora into two groups: bears and non-bears, while other three methods make more errors with the order of Carnivora. Moreover, only our distribution vector method puts pig in to the order of Artiodactyla successfully. In general, all the four methods can do the clustering with the animals and viruses corresponding to the evolution relationship. But our distribution vector method obtains the best results in the clustering.

In order to compare the speed of our method and the other three methods, we do the test on two sets of sequences. The first set consists of 8 datasets. The number of sequences in each dataset is 10, 20, 30, 40, 50, 60, 70 an 80 respectively where the lengths of all the sequences are around 4000. Another set also consists of 8 datasets. All the 8 datsets include 40 sequences. The lengths of all sequences in the 8 datasets are around 1000, 2000, 3000, 4000, 5000, 6000, 7000 and 8000 respectively. We build the phylogenetic tree on each dataset of the two sets by the four methods and record the time each method takes. The results in Figure 18 and Figure 19 show that our method is much faster than the other three methods. The time of our method increases linearly when the number of sequences or the length of sequences increases, whereas the acceleration of the time for the other three methods is much higher. The actual time differences are much higher than the visual differences in the figure since we are using the log(time) as the label of y-axis.

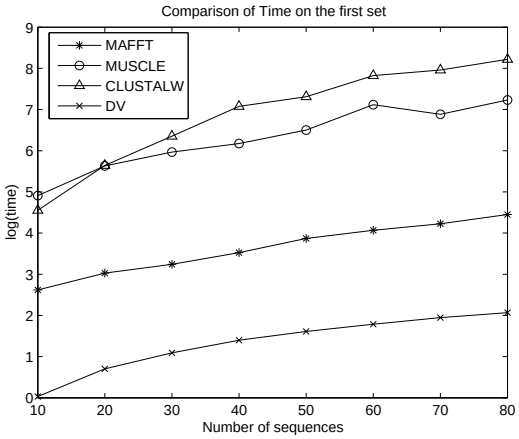

Figure 18. The Time Comparison of Four Methods on First Set

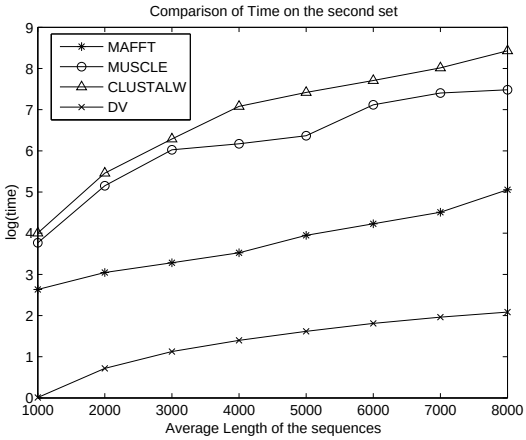

Figure 19. The Time Comparison of Four Methods on Second Set

Conclusion

This chapter introduces the distribution vectors to map the DNA sequences into the sixty dimensional Euclidean space. We prove that expectation and standard deviation of the distribution vectors do not depend on the length of the sequences. The experiments on the human DNA sequences and random sequences confirm the result. The determined expectation and standard deviation show that the distribution vector mapping is bounded and stable. Each component of the distribution vectors represents the distribution of one kind of nucleotide in k segments of the DNA sequence and plays the same important role in the mapping and clustering. Furthermore, we do the clustering on 80 mitochondrial complete genomes and the gene Haemagglutinin (HA) of 60 H1N1 viruses with our distribution vector method and other three methods. The phylogenetic trees we obtain show that the distances between the distribution vectors correspond to the evolutionary relationships between these sequences. Our method works for a set of genome sequences or a set of gene sequences. Most importantly, the distribution vector method is much faster than the other methods. Hence our method is more efficient to deal with huge datasets than the other methods. Especially, Our distribution vector method only needs to compute the distribution vector of a new sequence when it is put in the dataset, while those multiple sequence alignment methods have to do the multiple sequence alignment on the new dataset when a new sequence is added. It will be more practical to find the closest sequence to the new sequence in a huge dataset with our distribution vector method. Our method may help to discover the functionality or the evolution of the new sequence.

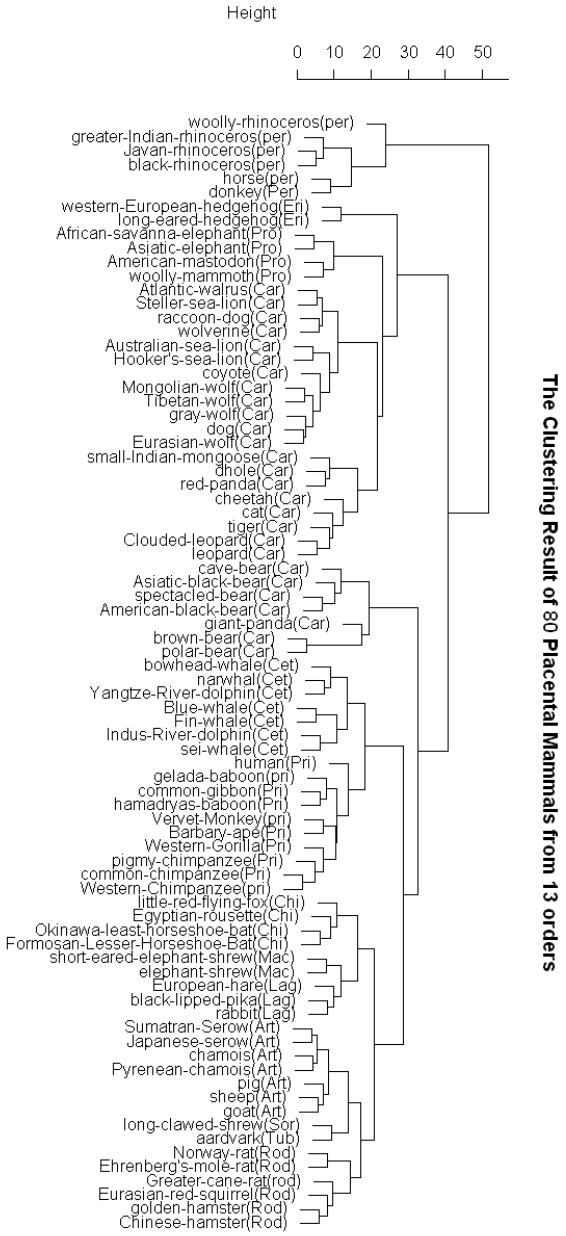

Figure 20. The Clustering result of 80 Mitochondrial Genomes

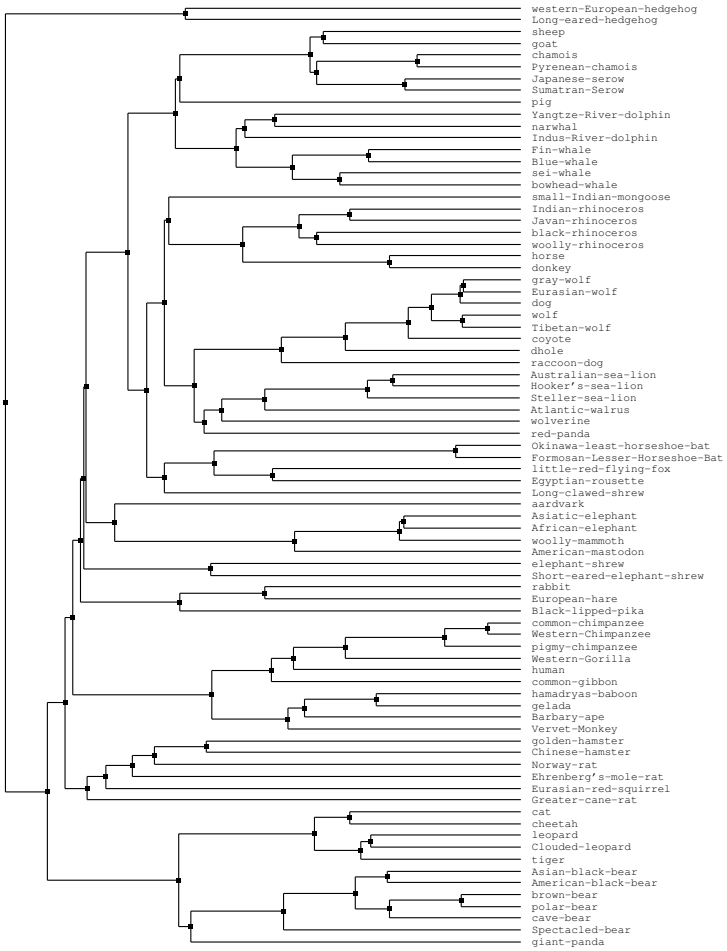

Figure 21. The Clustering result of 80 Mitochondrial Genomes by Clustalw

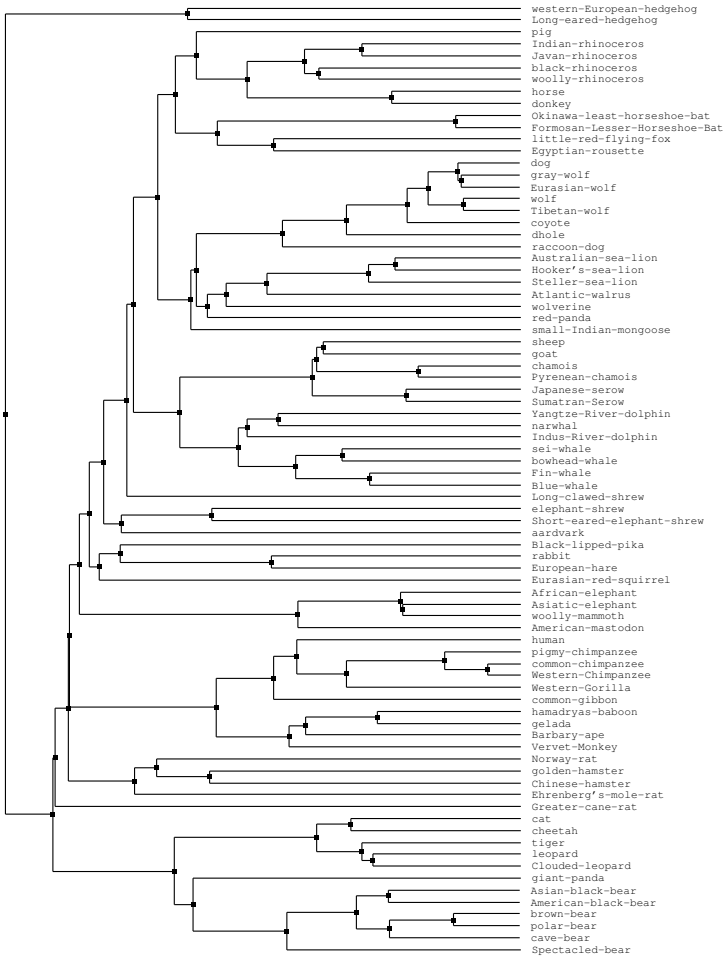

Figure 22. The Clustering result of 80 Mitochondrial Genomes by MAFFT

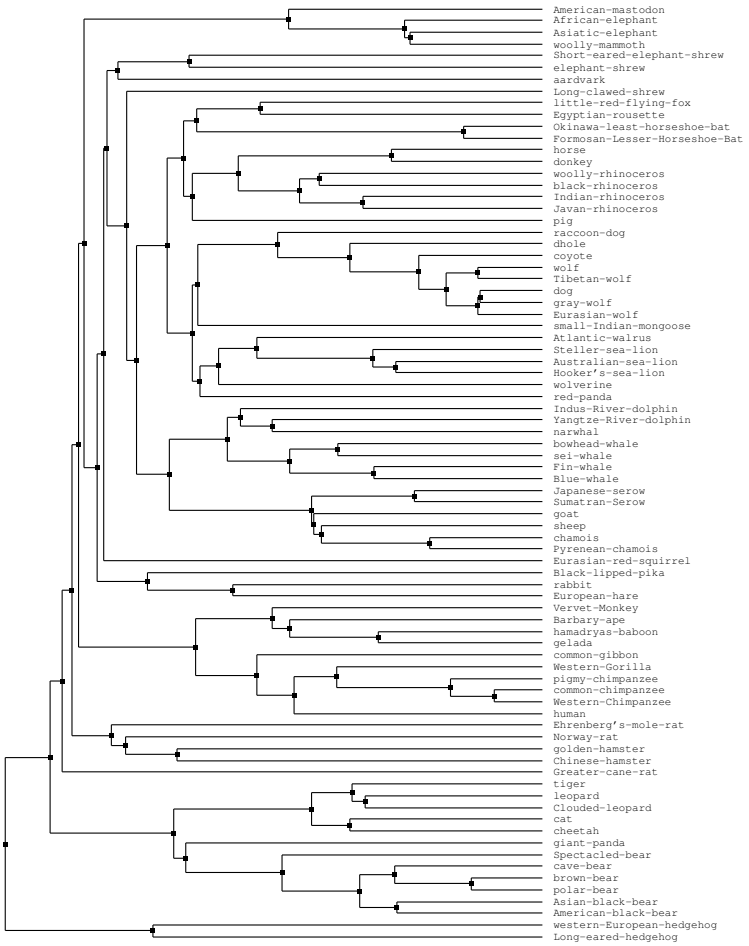

Figure 23. The Clustering result of 80 Mitochondrial Genomes by Muscle

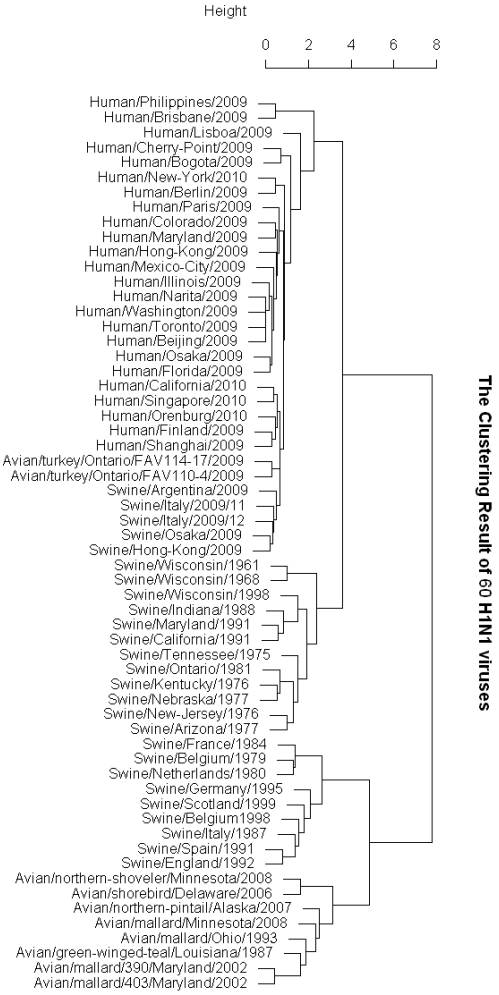

Figure 24. The Clustering result of 60 H1N1 viruses

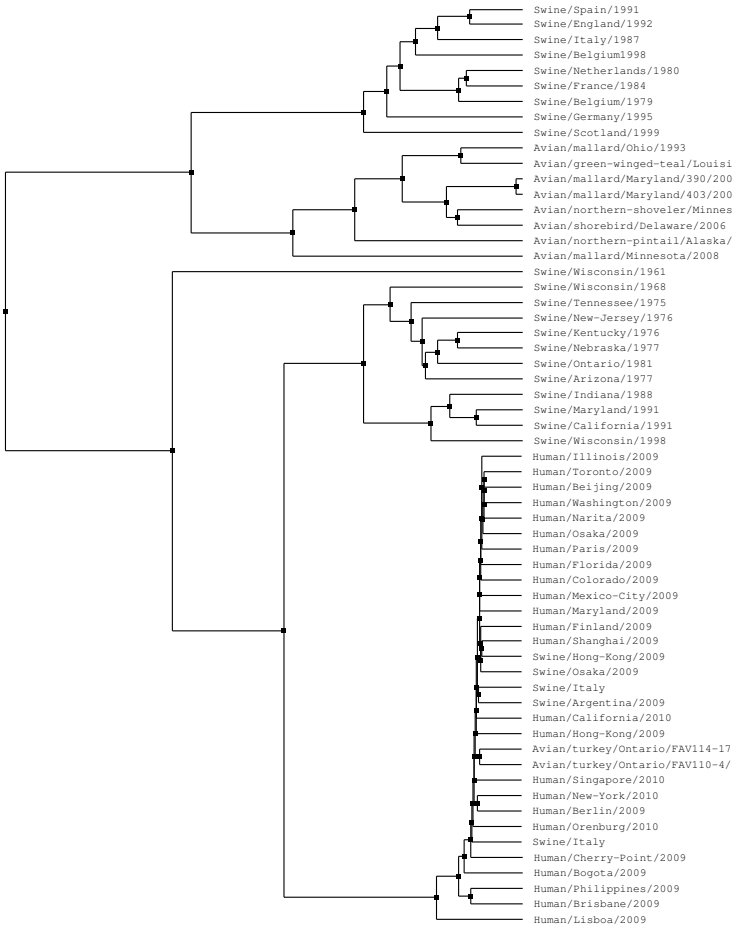

Figure 25. The Clustering result of 60 H1N1 viruses by Clustalw

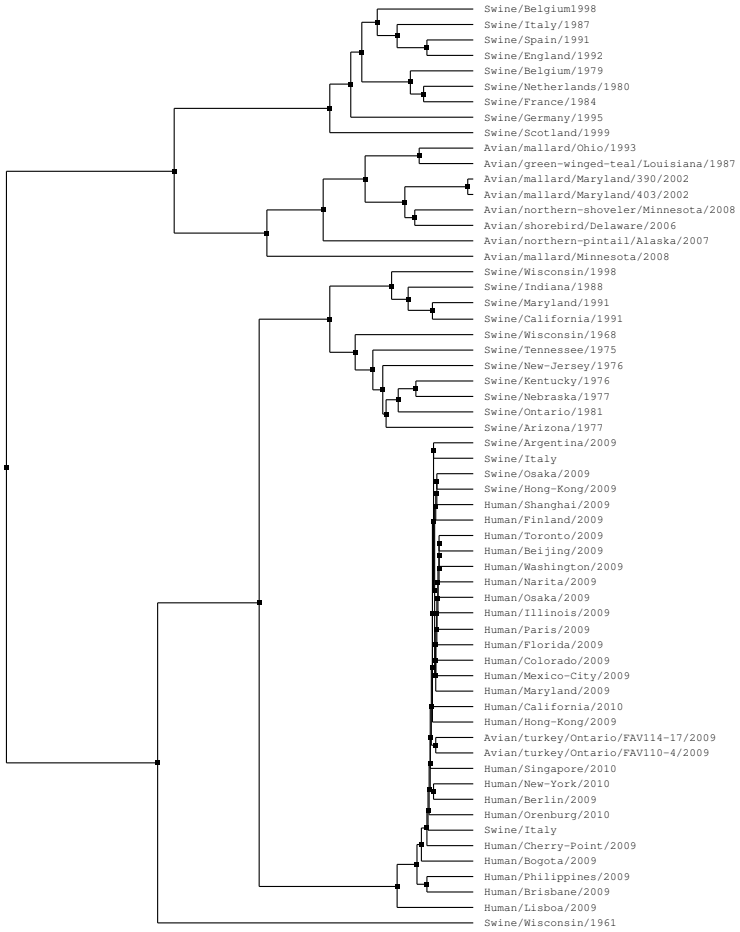

Figure 26. The Clustering result of 60 H1N1 viruses by MAFFT

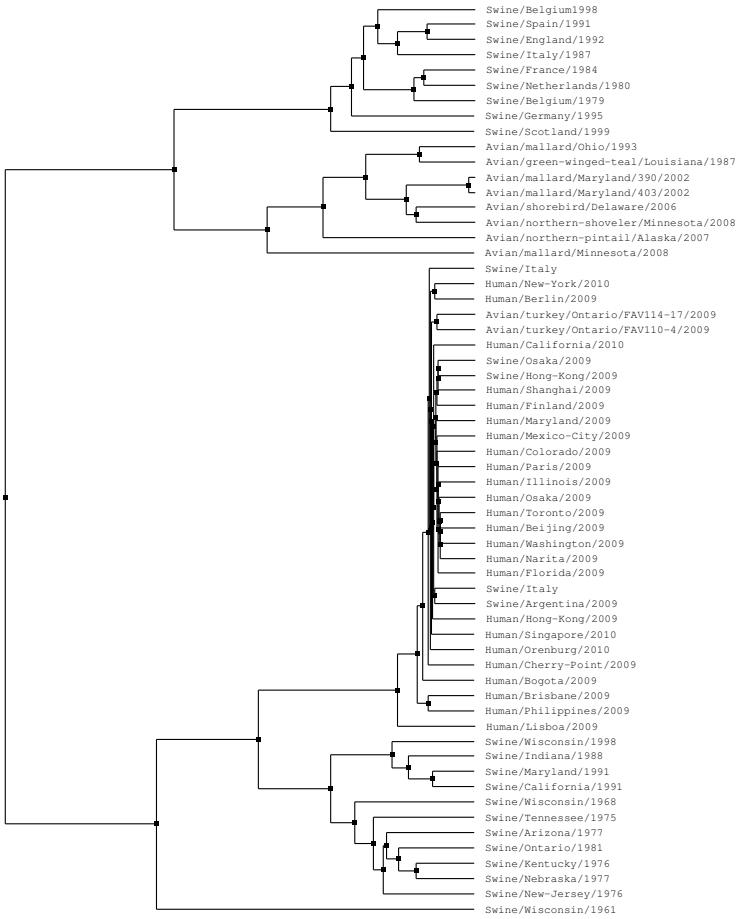

Figure 27. The Clustering result of 60 H1N1 viruses by Muscle

APPENDICES

TABLE IV: THE LIST OF 60 H1N1 VIRUSES

Source	**Location**	**Year**	**Accession**
Swine	Wisconsin	1998	AAF87282
	Belgium	1998	ACN67524
	Scotland	1999	ACO25069
	Argentina	2009	ADC32526
	Osaka	2009	BAI49135
	Italy	2009	ADA70669
	Hong Kong	2009	ADG08380
	Italy	2009	ADD84723
	Belgium	1979	ACO24983
	Netherlands	1980	AAD25309
	France	1984	ACO25089
	Italy	1987	AAD25310
	Spain	1991	ACO25122
	England	1992	ACO25133
	Germany	1995	CAP49183
	Wisconsin	1961	AAD25302
	Wisconsin	1968	ABV25636

Continued on next page

Table IV – continued from previous page

Source	**Location**	**Year**	**Accession**
Swine	Tennessee	1975	ABR28680
	New Jersey	1976	AAB39851
	Kentucky	1976	ABR28614
	Nebraska	1977	ABR28647
	Arizona	1977	ABU80287
	Ontario	1981	ABR28658
	Indiana	1988	ABF71860
	Maryland	1991	ABR29565
	California	1991	ABY84684
Mallard	Ohio	1993	ABM21960
	Maryland	2002	ABS70389
	Maryland	2002	ABS70400
	Minnesota	2008	ACT84288
Northern Shoveler	Minnesota	2008	ACT84833
Shorebird	Delaware	2006	ACU15899
Northern Pintail	Alaska	2007	ACY67472
Green Winged Teal	Louisiana	1987	ACZ48419

Continued on next page

Table IV – continued from previous page

Source	**Location**	**Year**	**Accession**
Turkey	Ontario	2009	ADI52835
	Ontario	2009	ADI52836
Human	Cherry Point	2009	ACY77544
	Bogota	2009	ACY77554
	Toronto	2009	ACQ44556
	Illinois	2009	ACS72651
	Colorado	2009	ACR49290
	Finland	2009	ACS50088
	Philippines	2009	ACR78158
	Shanghai	2009	ACR54974
	Osaka	2009	ACR46991
	Paris	2009	ACR43939
	Beijing	2009	ACR32998
	New York	2010	ADI99560
	California	2010	ADI99550
	Orenburg	2010	ADI99498
	Mexico City	2009	ADI49787

Continued on next page

Table IV – continued from previous page

Source	**Location**	**Year**	**Accession**
Human	Singapore	2010	ADI24597
	Berlin	2009	ADI49382
	Florida	2009	ACR08526
	Brisbane	2009	ACR08498
	Maryland	2009	ACR08538
	Washington	2009	ACR08543
	Narita	2009	ACR09395
	Lisboa	2009	ACR15748
	Hong Kong	2009	ACR18920

TABLE V: THE LIST OF 80 MITOCHONDRIAL GENOMES

Order	**Species**	**Accession**
Primates	Human	V00662
	Pigmy Chimpanzee	D38116
	Common Chimpanzee	D38113
	Western Gorilla	D38114
	Common Gibbon	X99256
	Hamadryas Baboon	Y18001
	Western Chimpanzee	GU112744
	Vervet Monkey	EF597501
	Gelada Baboon	FJ785426
	Barbary Ape	NC_002764
Proboscidea	African Elephant	AJ224821
	Asiatic Elephant	DQ316068
	American Mastodon	NC_009574
	Woolly Mammoth	NC_007596
Perissodactyla	Indian Rhinoceros	X97336
	Black Rhinoceros	NC_012682

Continued on next page

Table V – continued from previous page

Order	**Species**	**Accession**
Perissodactyla	Javan Rhinoceros	NC_012683
	Horse	NC_001640
	Donkey	NC_001788
	Woolly Rhinoceros	NC_012681
Macroscelidea	Elephant Shrew	AB096867
	Short-eared Elephant Shrew	NC_004026
Erinaceomorpha	Western European Hedgehog	X88898
	Long-eared Hedgehog	NC_005033
Rodentia	Greater Cane Eat	NC_002658
	Ehrenberg's Mole Rat	NC_005315
	Golden Hamster	NC_013276
	Norway Rat	X14848
	Chinese hamster	EU660217
	Eurasian Red Squirrel	AJ238588
Lagomorpha	Black-lipped Pika	NC_011029
	Rabbit	AJ001588
	European Hare	NC_004028
Artiodactyla	Pig	AJ002189

Continued on next page

Table V – continued from previous page

Order	**Species**	**Accession**
Artiodactyla	Sheep	AF010406
	Goat	AF533441
	Chamois	FJ207539
	Pyrenean Chamois	FJ207538
	Japanese Serow	NC_012096
	Sumatran-Serow	FJ207534
Cetacea	Yangtze River dolphin	NC_007629
	Sei Whale	NC_006929
	Narwhal	AJ554062
	Indus River Dolphin	NC_005275
	Fin Whale	NC_001321
	Blue Whale	NC_001601
	Bowhead Whale	AJ554051
Chiroptera	Okinawa Least Horseshoe Bat	NC_005434
	Little Red Flying Fox	NC_002619
	Egyptian Rousette	NC_007393
	Formosan Lesser Horseshoe Bat	NC_005433

Continued on next page

Table V – continued from previous page

Order	**Species**	**Accession**
Tubulidentata	Aardvark	Y18475
Carnivora	Tiger	EF551003
	Leopard	EF551002
	Clouded Leopard	NC_008450
	Cheetah	NC_005212
	Domestic Cat	U20753
	Tibetan Wolf	NC_011218
	Giant Panda	EF212882
	Asian Black Bear	DQ402478
	Brown Bear	AF303110
	Polar Bear	AF303111
	Spectacled Bear	EF196665
	Australian Sea Lion	NC_008419
	Hooker's Sea Lion	NC_008418
	Wolverine	NC_009685
	Dog	U96639
	Coyote	DQ480511

Continued on next page

Table V – continued from previous page

Order	**Species**	**Accession**
Carnivora	Gray Wolf	DQ480508
	American Black Bear	AF303109
	Raccoon Dog	NC_013700
	Dhole	GU063864
	Eurasian Wolf	NC_009686
	Red Panda	NC_009691
	Atlantic Walrus	NC_004029
	Steller Sea Lion	NC_004030
	Cave Bear	NC_011112
	Small Indian Mongoose	NC_006835
	Mongolian Wolf	EU442884
Soricomorpha	Long-clawed shrew	AB061527

CITED LITERATURE

1. Batzoglou, S. and Pachter, L.: Human and mouse gene structure: Comparative analysis and application to exon prediction. Genome Research, 10(7):950–958, July 2000.

2. Burge, C. and Karlin, S.: Prediction of complete gene structures in human genomic dna. Journal of Molecular Biology, 268:78–94, 1997.

3. Gross, S. and Brent, M.: Using multiple alignments to improve gene prediction. Journal of Computational Biology, 13(2):379–393, 2006.

4. Genetic Sequence Data Bank: The complete release notes for the current version of genbank. National Center for Biotechnology Information, 2010.

5. Fickett, J.: Recognition of protein coding regions in dna sequences. Nucleic Acids Research, 10(17):5303-5318, September 1982.

6. Burset, M. and Guigó, R.: Evaluation of gene structure prediction programs. Genomics, 34(3):353-367, 1996.

7. Blanco, E., Parra, G. and Guigó, R.: Using geneid to Identify Genes. Current Protocols in Bioinformatics, 2002.

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

9. Darwin C.: On the Origin of Species. John Murray, 1859.

10. Woese, C. and Fox, G.: Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences, 74(11):5088-5090, 1977.

11. Larkin M., Blackshields, G. and Brown, .N.: Clustal w and clustal x version 2.0. Bioinformatics, 23(21):2947–8, Nov 2007.

12. Edgar, R.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research., 32(5):1792–1797, 2004.

13. Katoh, K., Asimenos, G. and Toh, H.: Multiple alignment of dna sequences with mafft. Methods in Molecular Biology, 537:39–64, 2009.

14. Gate, M.: Simpler DNA sequence representations. Nature, 316(6025):219, Jul 18-24 1985.

15. Yau, S., Wang, J. and Niknejad, A.: Dna sequence representation without degeneracy. Nucleic Acids Research, 31(12):3078–3080, 2003.

16. Anastassiou, D.: Genomic signal processing. IEEE Signal Processing Magazine, 18(4):8–20, 2001.

17. Afreixo, V., Ferreira, P. and Santos, D.: Spectrum and symbol distribution of nucleotide sequence. Physical Review E, 70(3):031910, September 2004.

18. R Development Core Team: R: A Language and Environment for Statistical Computing. Vienna Austria R Foundation for Statistical Computing, 2009. ISBN 3900051070.

19. Coward, E.: Equivalence of two fourier methods for biological sequences. Journal of Mathematical Biology, 36(1):64 – 70, November 1997.

20. Flicek, P.: Gene prediction: compare and contrast. Genome Biology, 8(12), 2007.

21. Yin, C. and Yau, S.: Prediction of protein coding regions by the 3-base periodicity analysis of a dna sequence. Journal of Theoretical Biology, 247(4):687 – 694, 2007.

22. Lukashin, A. and Borodovsky, M.: GeneMark.hmm: New solutions for gene finding. Nucleic Acids Research, 26(4):1107–1115, 1998.

23. Parra, G., Agarwal, P., Abril, J., Wiehe, T., Fickett, J. and Guigó, R.: Comparative gene prediction in human and mouse. Genome Research, 13(1):108–117, January 2003.

24. Frenkel, F. and Korotkov, E.: Using Triplet Periodicity of Nucleotide Sequences for Finding Potential Reading Frame Shifts in Genes. DNA Research, 16(2):105–114, 2009.

25. Pinho, A., Neves, A., Afreixo, V., Bastos, C. and Ferreira, P.: A three-state model for dna protein-coding regions. Biomedical Engineering, IEEE Transactions on, 53(11):2148 –2155, nov. 2006.

26. Jiang, R. and Yan, H.: Segmentation of short human exons based on spectral features of double curves. International Journal of Data Mining and Bioinformatics, 2(1):15–35, 2008.

27. Hsieh, S., Chung, Y., Tang, C. and Lin, C.: Comparative exon prediction based on heuristic coding region alignment. Parallel Architectures, Algorithms, and Networks, International Symposium on, 0:14–19, 2005.

28. Alexandersson, M., Cawley, S. and Pachter, L.: SLAM: Cross-Species Gene Finding and Alignment with a Generalized Pair Hidden Markov Model. Genome Research, 13(3):496–502, 2003.

29. Beer, M. and Tavazoie, S.: Predicting gene expression from sequence. Cell, 117(2):185 – 198, 2004.

30. Yau, S., Yu, C. and He, R.: A protein map and its application. DNA and Cell Biology, 27(5):241–250, May 2008.

31. Waterhouse, A., Procter, J., Martin, D., Clamp, M. and Barton, G.: Jalview version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics, 25(9):1189–1191, May 2009.

32. Yu, C., Liang, Q., Yin, C., He, R. and Yau, S.: A Novel Construction of Genome Space with Biological Geometry. DNA Research, 17(3):155-168, 2010.

33. Raina, S., Faith, J., Disotell, T., Seligmann, H., Stewart, C. and Pollock, D.: Evolution of base-substitution gradients in primate mitochondrial genomes. Genome Research, 15(5):665–673, 2005.

34. Smith, G., Vijaykrishna, D., Bahl, J., Lycett, S., Worobey, M., Pybus,O., Ma, S., Cheung, C., Raghwani, J., Bhatt, S., Peiris, J., Guan, Y. and Rambaut, A.: Origins and evolutionary genomics of the 2009 swine-origin h1n1 influenza a epidemic. Nature, 459(7250):1122–1125, June 2009.

35. Solovyov, A., Palacios, G., Briese, T., Lipkin, W. and Rabadan, R.: Cluster analysis of the origins of the new influenza a(h1n1) virus. Eurosurveillance, 14, 2009.

36. Gorman, O., Bean, W., Kawaoka, Y., Donatelli, I., Guo, Y. and Webster, R.: Evolution of influenza A virus nucleoprotein genes: implications for the origins of H1N1 human and classical swine viruses. Journal of Virology, 65(7):3704–3714, 1991.

37. Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T.: Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proceedings of the National Academy of Sciences of the United States of America, 86(23):9355–9359, 1989.

38. Fitch, W. and Margoliash, E.: Construction of phylogenetic trees. Science, 155(760):279–84, Jan 1967.

39. Boore, J. and Brown, W.: Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. Current Opinion in Genetics and Development, 8(6):668–674, 1998.

40. Wolstenholme, D.: Animal mitochondrial DNA: Structure and evolution. International Review of Cytology, 141:173 – 216, 1992.

VITA

Contact

Bo Zhao

322 Science and Engineering Offices (M/C 249)

851 S. Morgan Street, Chicago, IL 60607

keanuzb@gmail.com

Education

- PH.D in Applied Mathematics, University of Illinois at Chicago. 2010
- M.S. in Applied Mathematics, University of Illinois at Chicago. 2006
- M.S. in Applied Mathematics, Xi'an Jiaotong University, China. 2002
- B.S. in Information & Computing Science, Xi'an Jiaotong University, China. 1999

Work Experience

- Computer Support (Graduate Assistant) at University of Illinois at Chicago

  Chicago,IL,USA, 2005-2010

- R&D Division Software Engineer, DaTang Telecom Technology Co.,Ltd

  Xi'an, ShaanXi, China, 2003-2004

Publications

1. Using Distribution Vectors to cluster homologous DNA sequences (with Stephen Yau and Rong He), 16pp. in ms., submitted for publication.

2. A Fast and Straightforward Method to Predict the Coding Regions on DNA sequences (with Stephen Yau), 10pp. in ms., submitted for publication.

3. A Novel Clustering Method via Nucleotide-Based Fourier Power Spectrum Analysis (with Victor Duan and Stephen Yau), 14pp. in ms., submitted for publication.