Binary Mapping and Genes Detection

1st Shikang Zhang
dept. Shanghai Jiao Tong University
zsk.11@sjtu.edu.cn

1st Yabo Tian
dept. Shanghai Jiao Tong University
a979376748@sjtu.edu.cn

1st Weilin Lang
dept. Shanghai Jiao Tong University
lwl000@sjtu.edu.cn

Abstract—By assigning proper(complex, in general) numerical values to each character, digital signal processing of biomolecular sequences provides a set of novel and useful tool. This article introduces the modeling process of the binary mapping scheme and evaluates the experimental results of it by comparing with some other existing mapping schemes. Digital signal processing of biomolecular sequences can be used to help related work about genes prediction. So we experiment to evaluate the effect of our binary mapping scheme in genes detection.

Index Terms—mapping, gene, prediction, spectrum, SNR

I. Introduction

THE main reason that the field of digital signal processing did not yet have significant impact on biomolecular sequence analysis is that the former refers to numerical sequences, while the latter refers to character strings. By assigning proper(complex, in general) numerical values to each character, digital signal processing of biomolecular sequences provides a set of novel and useful tool.

Identification of protein coding regions (exons) in eukaryotic genomic sequences is an active area of research at present. Mapping of symbolic genomic sequences to numeric sequences is the first step required for processing them using digital signal processing (DSP) tools. For DFT-based methods paired numeric and frequency of nucleotide are reported as the best mapping schemes. In this work performance of a waveletbased method for exon detection is evaluated with different symbolic-to-numeric representations.

With the development of genome sequencing for many organisms, more and more raw sequences need to be annotated. Gene prediction by computational methods for finding the location of protein coding regions is one of the essential issues in bioinformatics.

Since the beginning of the Human Genome Program (HGP) in 1990, databases of human and model organism DNA sequences have been increasing quickly. Computational gene prediction is becoming more and more essential for the automatic analysis and annotation of large uncharacterized genomic sequences. In the past two decades, many gene prediction programs have been developed.

With the exponential growth of genomic sequences, there is an increasing demand to accurately identify protein coding regions (exons) from genomic sequences. Despite many progresses being made in the identification of protein coding regions by computational methods during the last two decades, the performances and efficiencies of the prediction methods still need to be improved. In addition, it is indispensable to develop different prediction methods since combining different methods may greatly improve the prediction accuracy.

We analyzed the advantages and disadvantages of commonly used mapping schemes, and built a new mapping scheme with a fast computing speed. In order to test the actual effect of our mapping scheme, we do a comparative experiment with Voss mapping, Z-curve mapping and real number mapping schemes. The experimental results show that we can achieve as good experimental results as Voss mapping while computing faster

II. Related Work

For a very long DNA sequence, when calculating its power spectrum or signal-to-noise ratio, the overall calculation amount of the discrete Fourier transform (DFT) is still very large, which will affect the efficiency of the designed gene recognition algorithm.

When calculating the power spectrum and signal noise ratio of long sequences, the process of discrete fourier transform is a big problem. Many mapping schemes have been widely used for this problem. Among them, Voss mapping, tetrahedron mapping, Z-curve mapping and real number mapping are the main ones.

Mapping Schemes	DNA Representation	S(n)= [CGAT]	Complexity
Voss	$X_n = 1$ for $S(n) = x$ $X_n = 0$ for $S(x) \neq x$ X_n applies to any C_n, G_n, A_n, T_n	$C_n = [1, 0, 0, 0]$ $G_n = [0, 1, 0, 0]$ $A_n = [0, 0, 1, 0]$ $T_n = [0, 0, 0, 1]$	4
Tetrahedron	$x_r(n) = \frac{\sqrt{2}}{3}[2T_n - C_n - G_n]$ $x_g(n) = \frac{\sqrt{6}}{3}[C_n - G_n]$ $x_b(n) = \frac{1}{3}[3A_n - T_n - C_n - G_n]$	$x_r(n) = \frac{\sqrt{2}}{3}[-1,-1,0,2]$ $x_g(n) = \frac{\sqrt{6}}{3}[-1,-1,0,0]$ $x_b(n) = \frac{1}{3}[-1,-1,3,-1]$	3
Z-Curve	$\begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} = 2 * \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} X_A[n] \\ X_C[n] \\ X_G[n] \\ X_T[n] \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} = \begin{bmatrix} -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ -1 & -1 & 1 & 1 \end{bmatrix}$	3
Real Number	A=0, G=1, C=2, T=3	[2, 1, 0, 3]	1

Fig. 1. Different Mapping Schemes Comparison

Voss mapping is a basic mapping scheme and tetrahedron mapping is a kind of dimension reduction of voss

mapping. Z-curve mapping is a magical method which has excellent biological significance. And the real number mapping is a method that strives for computing efficiency but ignores part of practical value.

III. Modeling

A. Power Spectrum and SNR

1) A New Mapping Scheme:

In order to digitize the DNA , we need use a sequences to expresses the DNA, we Let I = A, T, G, C, any DNA sequence of length N can be expressed as:

$$S = \{S[n] \mid S[n] \in I, n = 0, 1, ..., N-1\} \tag{1}$$

That is, the symbol sequence S of A, C, T and G: S[0], S[1],..., S[n-1]. For any given n less than N, we use two sequences to record the n-th base. Concretely speaking, we use a[n]=0, b[n]=0 to express S[n]=A, we use a[n]=1, b[n]=0 to express S[n]=C, we use a[n]=0, b[n]=1 to express S[n]=T, and we use a[n]=1, b[n]=1 to express S[n]=G. i.e.

$$a[n] = \begin{cases} 1 & \text{S[n]=C/G} \\ 0 & \text{S[n]=A/T} \end{cases}, \quad b[n] = \begin{cases} 1 & \text{S[n]=T/G} \\ 0 & \text{S[n]=A/C} \end{cases} \tag{2}$$

2) Fourier Transform Analysis:

In order to study the characteristics of DNA coding sequences. Discrete Fourier transform(DFT) should be performed on coding sequences. But it's too complex to perform DFT on string. So, we can map DNA coding sequences to two 0-1 sequences through our mapping.

Assuming that a given DNA sequence fragment is S=ACTGAGCT, the two generated 0-1 sequences can be divided into two groups using our mapping shown in euqation 2:

$$\{a[n]\} : \{0, 1, 0, 1, 0, 1, 1, 0\}$$
$$\{b[n]\} : \{0, 0, 1, 1, 0, 1, 0, 1\}$$

For these two 0-1 sequences, we can easily perform DFT of them through DFT formula:

$$\begin{aligned} A[k] &= \sum_{n=0}^{N-1} a[n] e^{-j\frac{2\pi nk}{N}}, \ k = 0, 1, ..., N-1 \\ B[k] &= \sum_{n=0}^{N-1} b[n] e^{-j\frac{2\pi nk}{N}}, \ k = 0, 1, ..., N-1 \end{aligned} \tag{3}$$

3) Calculating Power Spectrum and SNR:

The square power spectrum of the complex sequence P[k] and Q[k] was calculated and added to obtain the spectrum power sequence {P[k]} of the whole DNA sequence S:

$$P[k] = |A[k]|^2 + |B[k]|^2 \tag{4}$$

The power spectrum curve of the exon sequence has a large peak at the frequency of k=N/3, while the intron has no similar peak. This statistical phenomenon is called the "triple periodicity of bases".

The average value of the total power of DNA sequence S is:

$$\bar{E} = \frac{\sum_{k=0}^{N-1} P[k]}{N} \tag{5}$$

The ratio of the power spectrum value of DNA sequence at a specific position, i.e. k=N/3, to the average value of the total power spectrum of the whole sequence s is taken as the "signal-noise ratio"(SNR) of the DNA sequence:

$$R = \frac{P[\frac{N}{3}]}{\bar{E}} \tag{6}$$

The SNR of a DNA sequence not only indicates the relative height of Peak Value, but also reflects the strength of triple periodicity of coding or non coding sequence.

The exons of DNA sequences usually satisfy that the SNR is higher than a properly selected threshold (such as 2), while introns generally will not have this property.

B. Genes Detection

As the SNR of the exon and the intron is obviously different, we can detect and predict all exons of a complete DNA sequence that has not been annotated by using this characteristic. At present, there are two kinds of gene recognition algorithms based on SNR, "Fixed length sliding window method" and "Moving sequence method". We applied our mapping scheme to these two widely used algorithms

1) Fixed length sliding window method:

For a DNA sequence s and its indicator sequences a[n] and b[n], n=0,1,2,...,N-1. Take the length M (usually taken as a multiple of 3, such as M = 99, 129, 255, 513, etc.) as the fixed window length.

For any ($0 \le n \le N-1$), On the sequence fragment $[n - \frac{M-1}{2}, n + \frac{M+1}{2}]$ with length M centered on n (when n is close to both ends of the sequence, the actual effective length of the window may be less than M), do Discrete Fourier Transform (DFT) for 2 indicator sequences.

$$\begin{aligned} A[k] &= \sum_{i=n-\frac{M-1}{2}}^{n+\frac{M+1}{2}} a[i] e^{-j\frac{2\pi ik}{N}}, \ k = 0, 1, ..., M-1 \\ B[k] &= \sum_{i=n-\frac{M-1}{2}}^{n+\frac{M+1}{2}} b[i] e^{-j\frac{2\pi ik}{N}}, \ k = 0, 1, ..., M-1 \end{aligned} \tag{7}$$

The total spectrum $p(n; \frac{M}{3})$ at $\frac{M}{3}$ is shown as below:

$$P[\frac{M}{3}] = |A[\frac{M}{3}]|^2 + |B[\frac{M}{3}]|^2 \triangleq P(n; \frac{M}{3}) \tag{8}$$

Finally, just normalize the frequency spectrum value $p(n; \frac{M}{3})$, n=0,1,2,...,N-1.

2) Moving sequence method:

Let S be a known DNA sequence and its indicative sequence is a[n] and b[n], n=0,1,2,...N-1. For any $n(0 < n \leq N-1)$, n is usually a multiple of 3 and increases gradually. On a sequence fragment [0, n-1] with length n on the left of n, the corresponding subsequence of DNA sequence S is called the "mobile subsequence" of DNA sequence $S_{0\sim n-1}$. Do Discrete Fourier Transform (DFT) on two indicator sequences corresponding to the mobile sequence:

$$\begin{aligned} A[k] &= \sum_{i=0}^{n-1} a[i] e^{-j\frac{2\pi ik}{M}}, \ k = 0, 1, ..., M-1 \\ B[k] &= \sum_{i=0}^{n-1} b[i] e^{-j\frac{2\pi ik}{M}}, \ k = 0, 1, ..., M-1 \end{aligned} \tag{9}$$

Then, calculate the SNR of moving subsequence:

$$R[n] = \frac{P[\frac{n}{3}]}{\bar{E}[n]} = \frac{|A[\frac{n}{3}]|^2 + |B[\frac{n}{3}]|^2}{\bar{E}[n]}, \tag{10}$$

where $\bar{E}$ is the average power spectrum $\bar{E}[n] = \frac{\sum_{k=0}^{n-1} P[k]}{n}$ of the moving subsequence S.

IV. Results

After we built the model, we did some experiments to evaluate the effect of our model. We did the following experiments:

1.Power spectrum and signal-noise ratio experiment of voss mapping, z-curve mapping, real number mapping, binary mapping.

2.Compare and evaluate the experiment results of different mapping schemes.

3.Evaluate the experiment results in genes detection using binary mapping.

A. Power Spectrum and SNR

The following graphs is a comparison of different mapping schemes's results of a lot of experiments.

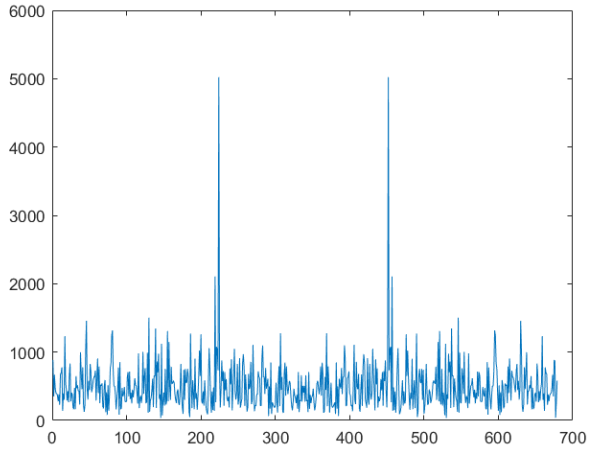

Fig. 2. Voss mapping Power Spectrum

This is the result power spectrum graph of Voss mapping scheme, and its SNR is 7.3421. It can be seen from the figure that the power spectrum of Voss mapping has an obvious peak at $\frac{N}{3}$ and $\frac{2N}{3}$, indicating that the scheme conforms to the triple periodicity of bases.

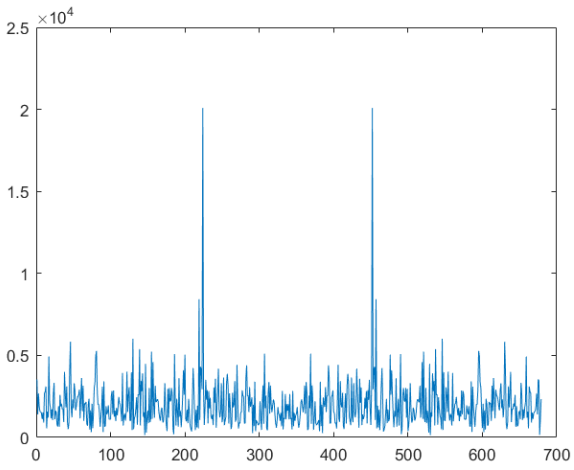

Fig. 3. Z-curve mapping Power Spectrum

This is the result power spectrum graph of Z-curve mapping scheme, and its SNR is 9.7895. It can be seen from the figure that the power spectrum of Z-curve mapping has an obvious peak at $\frac{N}{3}$ and $\frac{2N}{3}$, indicating that the scheme conforms to the triple periodicity of bases. And Z-curve mapping obviously has a higher SNR, and the effect is excellent, mainly because the mapping is proposed from the biological principle, so the processing of bases is better.

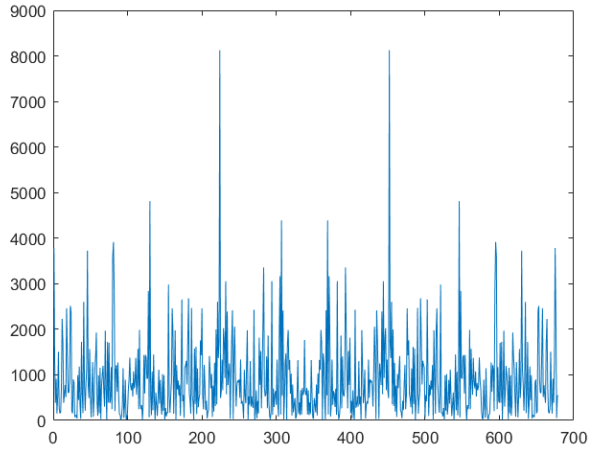

Fig. 4. Real mapping Power Spectrum

This is the result power spectrum graph of Real mapping scheme, and its SNR is 3.2422. It can be seen from the figure that the power spectrum of Real mapping has an obvious peak at $\frac{N}{3}$ and $\frac{2N}{3}$, indicating that the scheme also conforms to the triple periodicity of bases. However, it can be clearly seen from the figure that the SNR value is not high, indicating that the noise of the mapping is too large, and it cannot display the triple periodicity of the base as well as the previous two mappings.

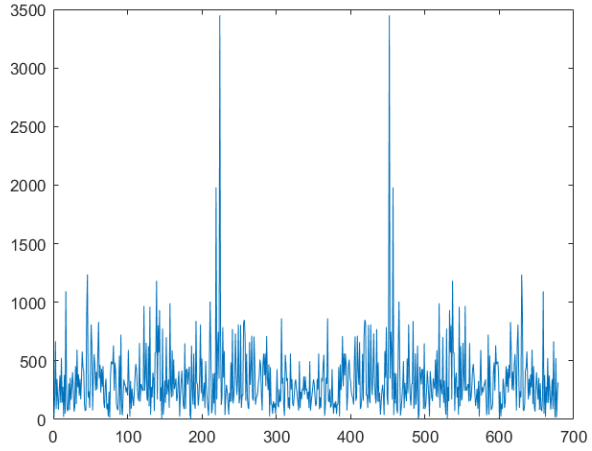

Fig. 5. Binary mapping Power Spectrum

This is the result power spectrum graph of Binary mapping scheme, and its SNR is 5.8458. It can be seen from the figure that the power spectrum of our Binary mapping has an obvious peak at $\frac{N}{3}$ and $\frac{2N}{3}$, indicating that the scheme conforms to the triple periodicity of bases. And this mapping has a lower SNR than the real mapping, the effect is better, and the base three periodicity performance is more obvious.

We implement the four schemes on 13 sequences of Human mtDNA, obtaining the following table that demonstrates our results, of which each row represents the performances of different mapping schemes on a particular sequence.

TABLE I
Experiment Results

SNR	zcurve	voss	real	binary
NC_012920_1_cds_Seq1	16.0930	12.0698	8.3001	10.6052
NC_012920_1_cds_Seq2	13.0857	9.8143	7.9654	8.2060
NC_012920_1_cds_Seq3	32.6468	24.4851	10.3275	20.8106
NC_012920_1_cds_Seq4	9.7895	7.3421	3.2422	5.8458
NC_012920_1_cds_Seq5	1.4193	1.0645	1.1427	1.3150
NC_012920_1_cds_Seq6	14.8370	11.1278	11.4896	7.1190
NC_012920_1_cds_Seq7	5.3624	4.0218	2.0417	3.6005
NC_012920_1_cds_Seq8	4.8821	3.6615	1.8931	3.8906
NC_012920_1_cds_Seq9	7.5241	5.6431	3.8811	4.8514
NC_012920_1_cds_Seq10	17.3332	12.9999	10.0151	11.6665
NC_012920_1_cds_Seq11	30.5401	22.9051	14.3293	20.0282
NC_012920_1_cds_Seq12	5.6000	4.2000	2.7081	1.1264
NC_012920_1_cds_Seq13	18.5787	13.9341	4.4837	15.5884

In terms of calculation, the main time to calculate the power spectrum and signal-noise ratio is spent on discrete fourier transform. Voss mapping requires four discrete fourier transforms, and Z-curve mapping requires three discrete fourier transforms, binary mapping requires two discrete fourier transforms, and real number mapping requires one discrete fourier transform. In terms of calculation speed, binary mapping is 2 times faster than Voss mapping and 1.5 times Z-curve mapping.

In terms of the final result, through analyzing the multiple sets of data obtained from our experiment, the real number mapping obtained signal-noise ratio is generally low, the experimental effect is not satisfactory. In most cases, it is worse a lot than binary mapping. The Z-curve mapping has the best experimental effect. And the results of binary mapping is as good as Voss mapping.

Although the real number mapping is the fastest mapping of the four mapping schemes, the signal-noise ratio of real number mapping is generally low. Most of DNA sequence fragments can't get a satisfactory result. Although the experimental effect of our binary mapping is not as perfect as Z-Curve, but our binary mapping can get the enough high SNR and the obvious triple periodicity with only $\frac{2}{3}$ time of Z-Curve mapping. Generally speaking, our binary mapping has such a good comprehensive effect.

B. Genes Detection

Our mapping method has achieved good results in calculation of power spectrum and SNR. Next, we use our binary mapping on the genes detection of all the sequences of Human mtDNA, and use two methods to evaluate our binary mapping scheme.

1) Fixed length sliding window method:

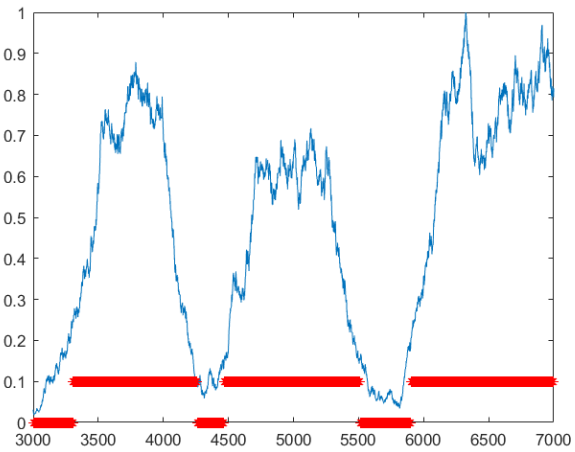

Fig. 6. Binary mapping

The upper red line in the figure represents the exon region, and the lower red line represents the intron region. As can be seen from the figure, the boundary between exons and introns is very obvious, and the two can be easily distinguished, indicating that our binary mapping has a good effect on gene recognition.

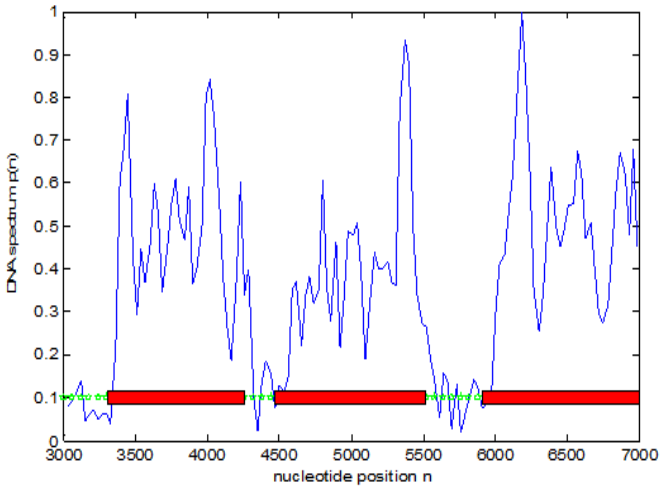

Fig. 7. Voss mapping

Similar to the above results, the red line in the figure represents the exon region, and can be seen that Voss mapping can also clearly distinguish the two.

By analyzing the above two results, we can find that the results obtained by the two mapping methods are very similar, and both have good recognition results, which shows that our binary mapping can use the Fixed length sliding window method smoothly.

2) Moving sequence method:

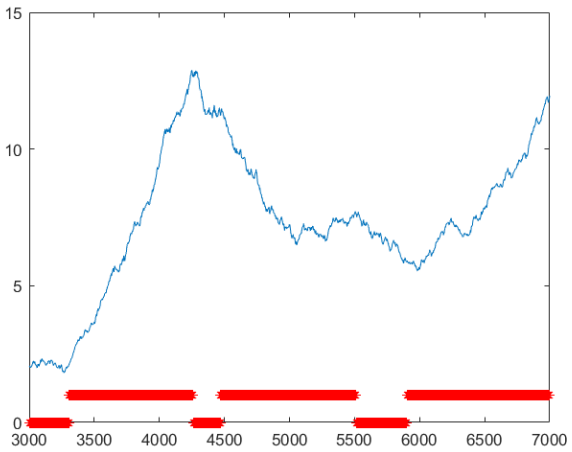

Fig. 8. Binary mapping

The upper red line in the figure represents the exon region, and the lower red line represents the intron region. As can be seen from the figure, We can relatively clearly distinguish two regions, indicating that our binary mapping has a good effect on gene recognition using Moving sequence method.

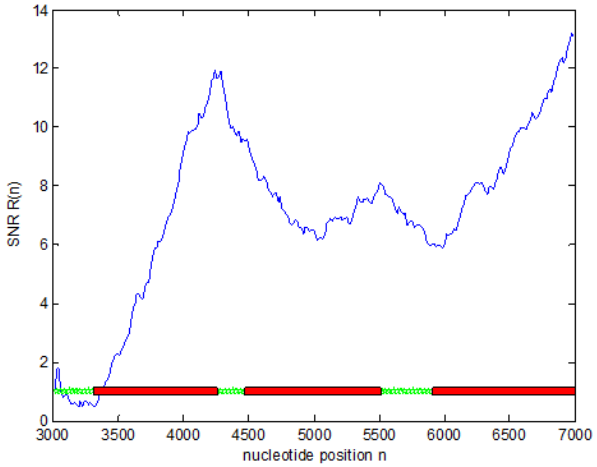

Fig. 9. Voss mapping

The red line in the figure represents the exon region, and the green line represents the intron region. Similar to the above results, Voss mapping can also relatively clearly distinguish two regions.

By analyzing the above two results, we can find that the results obtained by the two mapping methods are very similar, and both have good recognition results, which shows that our binary mapping can use the moving sequence method smoothly.

Binary mapping can well support fixed length sliding window algorithm and moving sequence algorithm in genes

detection. So we can use binary mapping to do genes detection and other related work.

V. Conclusion

In this article, we propose a new mapping scheme, which can map a DNA sequence into two numerical sequences, and then the sequence can be analyzed by the DSP method. By doing DFT on these two sequences, we can achieve the purpose of analyzing the original DNA sequence.

Our team analyzed the power spectrum of different mapping schemes. Through comparison with different mapping schemes, we found that the binary mapping we proposed can well meet a series of requirements for DNA sequence analysis. A complete set of DNA sequence analysis methods including gene detection methods can be applied to our proposed binary mapping. Moreover, our binary mapping can also obtain better speed and close recognition results than Voss mapping. In theory, our binary mapping is twice as fast as the voss mapping.

In conclusion, our binary mapping is a well-performing method considering both computational efficiency and experimental results, and provides more choices for gene prediction.

References

[1] Z. Wang, Y. Chen, and Y. Li, "A brief review of computational gene prediction methods," Genomics, proteomics & bioinformatics, vol. 2, no. 4, pp. 216–221, 2004.

[2] M. Yan, Z.-S. Lin, and C.-T. Zhang, "A new fourier transform approach for protein coding measure based on the format of the z curve." Bioinformatics (Oxford, England), vol. 14, no. 8, pp. 685–690, 1998.

[3] S. D. Sharma, K. Shakya, and S. Sharma, "Evaluation of dna mapping schemes for exon detection," in 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET). IEEE, 2011, pp. 71–74.

[4] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," Bioinformatics, vol. 16, no. 12, pp. 1073–1081, 2000.

[5] D. Kotlar and Y. Lavner, "Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions," Genome research, vol. 13, no. 8, pp. 1930–1937, 2003.

[6] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic dna," Journal of molecular biology, vol. 268, no. 1, pp. 78–94, 1997.

[7] C. Yin and S. S.-T. Yau, "Prediction of protein coding regions by the 3-base periodicity analysis of a dna sequence," Journal of theoretical biology, vol. 247, no. 4, pp. 687–694, 2007.

[8] M. Berryman, A. Allison, C. Wilkinson, and D. Abbott, "Review of signal processing in genetics," Fluctuation and Noise Letters, vol. 5, no. 04, pp. R13–R35, 2005.