A Novel Gene Detection Method Based on Period-3 Property

Lun Huang, Mohammad Al Bataineh, G. E. Atkin, *Senior Member, IEEE*, Siyun Wang, Wei Zhang

***Abstract*—Processing of biomolecular sequences using communication theory techniques provides powerful approaches for solving highly relevant problems in bioinformatics by properly mapping character strings into numerical sequences. We provide an optimized procedure for predicting protein-coding regions in DNA sequences based on the period-3 property of coding region. We present a digital correlating and filtering approach in the process of predicting these regions, and find out their locations by using the magnitude of the output sequence. These approaches result in improved computational techniques for the solution of useful problems in genomic information science and technology.**

I. Introduction

THE application of Fourier transform techniques to the research of DNAs and protein sequences results in a very important discovery that is the period-3 property in the protein-coding region. It is convenient to introduce indicator sequences for bases in DNA. For example the indicator for base A is a binary sequence of the form

$$x_A(n) = 100110101000101010...$$

where 1 indicates the presence of an A and 0 indicates its absence. The indicator sequences for the other three bases are defined similarly. Denote the Discrete Fourier Transform (DFT) of a length-N block of $x_A(n)$ as $X_A(k)$, that is

$$X_A(k) = \sum_{n=0}^{N-1} x_A(n) e^{-j2\pi kn/N}, 0 \le k \le N-1 \qquad (1)$$

The DFTs of $X_G(k)$, $X_T(k)$, and $X_C(k)$ are defined in the same way.

It has been noticed that protein-coding regions (exons) in genes have a period-3 component because of coding biases in the translation of codons into amino acids. This observation can be traced back to the research work of Trifonov and Sussman in 1980 [1]. The period-3 property is not present outside exons, thus can be exploited to locate exons. Taking N to be a multiple of 3 and plot

L. Huang, PhD Candidate; Mohammad Al Bataineh, Ph.D. Candidate; and G. E. Atkin, Ph. D., Senior Member IEEE, Associate Professor; are with the Department of Electrical & Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616; 1-312-567-3417; 1-312-567-8976 Fax (e-mail: lhuang13@iit.edu; albamoh@iit.edu; atkin@iit.edu).

Siyun Wang, PhD candidate, and Wei Zhang, Ph. D., Assistant Professor; are with the Department of Biological, Chemical, Physical Sciences, Illinois Institute of Technology, Chicago, IL 60616; 312-567-3123; 312-567-3494 fax (e-mail: swang26@iit.edu, zhangw@iit.edu).

$$S(k) = |X_A(k)|^2 + |X_T(k)|^2 + |X_C(k)|^2 + |X_G(k)|^2 \qquad (2)$$

Then we should observe a peak at the sample value k = N/3 (corresponding to 2π/3). Given a long sequence of bases we can calculate S(N/3) for short windows of the data, and then slide the window through the overall sequence. Thus, we can get a picture of how S(N/3) evolves along the DNA sequence. It is necessary that the window length N be sufficiently large (typical window sizes are a few hundreds to a few thousands) so that the periodicity effect dominates the background 1/f spectrum. However, a long window implies larger computation complexity in predicting the exon location.

It has been claimed that the period-3 property is due to non-uniform codon usage, also known as codon bias [2]; even though there are several codons which code a given amino acid, they are not used with uniform probability in organisms. For example, base G dominates at certain codon positions in the coding regions [3]. Experiments show that the use of the plot $|X_G(k)|^2$, which depends on base G alone, is often sufficient for revealing the period-3 property, and therefore for the detection of protein coding regions.

However, this method has a lot of limits. For example, it requires the pre-knowledge of the length of the coding region to perform DFT. In this paper, a novel detection approach is proposed. It uses correlation, maximum ratio combination, and filtering. This approach does not require any pre-knowledge of the coding regions, and simulation results show that it can effectively detect the profile of the reading frame for a given genome sequence. This paper is organized as follows. In section II, the system model and underlying theories are described; in section III, the simulation results are analyzed; section IV presents the conclusion of the proposed approach.

II. System Model and Theories

The system model for the proposed detection scheme is shown in Figure 1.

As we know, the genome sequence includes 4 types of bases, which are A, T, C, G. Before correlation, we must map these bases to symbols. In this paper, we use a complex poly-phase set {1, j, -1,-j}.

The mapping sequence of genome is correlated with four sequences. These four sequences are composed by using 3 out the 4 types of bases in a period-3 pattern. For example, ATCATC…., can be one of these sequences. Obviously, there are at least four sequences corresponding to 4 base combinations.

Assuming that the input complex genome sequence is

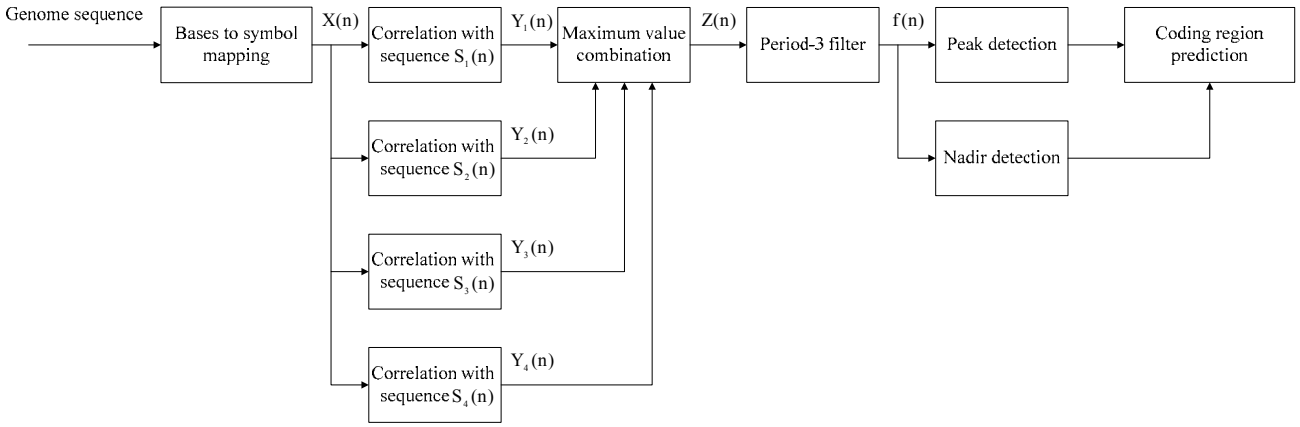

Figure 1. Block diagram for proposed protein coding region detection approach.

X(n), the four period-3 sequences are $S_1(n)$, $S_2(n)$, $S_3(n)$, $S_4(n)$, then the correlation outputs are:

$$Y_i(n) = X(n) \otimes S_i(-n) \tag{4}$$

Where '$\otimes$' denotes convolution. After correlation, the 4 output sequences are normalized respectively, and combined by maximum ratio combination algorithm.

$$Z(n) = \sum_{i=1}^{4} \left[Y_i(n) \cdot \frac{|Y_i(n)|}{\left| \sum_{j=1}^{4} Y_j(n) \right|} \right] \tag{5}$$

Period-3 filtering can be regarded as the sliding window method. A simple filter has an impulse response

$$w(n) = \begin{cases} e^{j\omega_0 n} & 0 \le n \le N-1 \\ 0 \quad otherwise \end{cases} \tag{3}$$

This is a bandpass filter with minimum stopband attenuation of about 13 dB. The passband should be centered at $\omega_0 = 2\pi/3$, thus, the period-3 filter is defined as

$$W(n) = \begin{cases} e^{j\cdot\frac{2\pi}{3}\cdot n} & 0 \le n \le N-1 \\ 0 \quad otherwise \end{cases} \tag{6}$$

This is a sequence of length N. Carefully designed digital filter can isolate the period-3 behavior from background 1/f noise more effectively. We can also use efficient methods to design and implement the filter to reduce computational complexity.

The output of period-3 filter is

$$f(n) = \sum_{i=0}^{N-1} Z(n-i)W(i)\text{, n = N-1, L, … , } L_x - 1. \tag{7}$$

Where L_x is the length of input complex genome sequence X(n).

In peak detection and coding region prediction, a window with length L (L is odd) is applied. When we slide this window on the filter output sequence f(n), at each time n = k, we can obtain a maximum value at each time n=k, and

$$Max(k) = \max\{f(k+i-L/2), \quad i = 1,2,...,L\}. \tag{8}$$

If $Max(k) = f(k)$, it can be judged that there is a peak at n=k.

It is well known that, for mRNA, synthesis of a peptide always starts from methionine (Met), coded by AUG. The stop codon (UAA, UAG or UGA) signals the end of a peptide. For DNA, U (uracil) should be replaced by T (thymine). In a DNA molecule, the sequence from an initiating codon (ATG) to a stop codon (TAA, TAG or TGA) is called an open reading frame (ORF), which is likely (but not always) to encode a protein or polypeptide. ORF can be combined with gene code books to improve its performance of gene detection.

Furthermore, to improve the coding regions detection knowledge of the non-coding regions can be used. There are many regulatory sequences in the non-coding regions, then it is possible that the non-coding region can be located with the help of these regulatory sequences in condition that enough large regulatory sequence set is available. Fake ORF could be eliminated by using this knowledge of non-coding region.

On the other hand, the approach proposed in this paper can be used to detect non-coding region with only slightly changes.

$$Min(k) = \min\{f(k+i-L/2), \quad i = 1,2,...,L\} \tag{9}$$

When $Min(k) = f(k)$, it can be decided that there is a nadir at n=k. Generally, the nadir locates in non-coding region, so the non-coding region can be found by using this approach.

What should be noted is that the proposed approach can detect the protein coding region in both 5'-3' and 3'-5' genome sequence. When a peak appears, it indicates that a coding region exists on this location of either 5'-3' or 3'-5' genome sequence, even both. While a nadir indicates a non-coding region for both 5'-3' and 3'-5' genome sequence.

III. SIMULATION RESULTS AND ANALYSIS

In simulation, a genome segment from prokaryotic bacteria E. coli strain MG1655 is used. This segment has 40386 nucleotides.

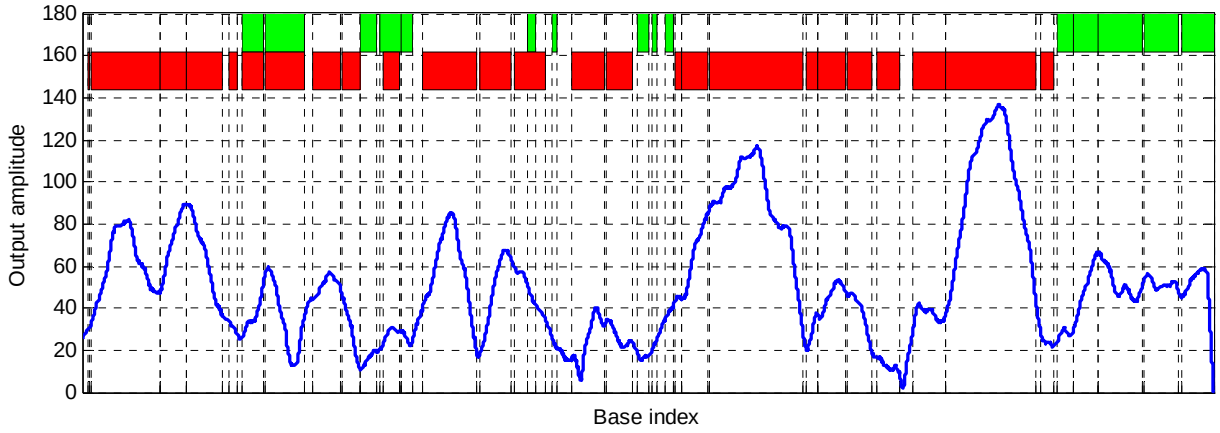

Figure 2. The absolute value of period-3 filter output

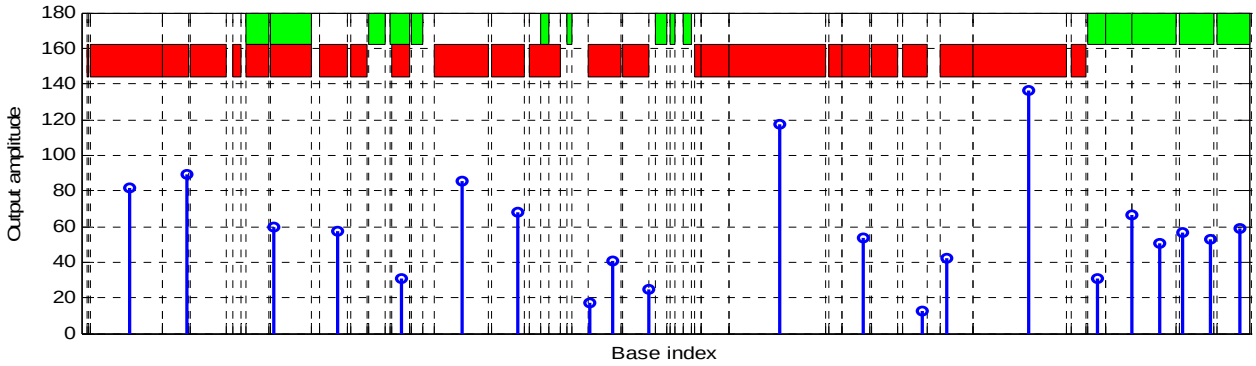

Figure 3. The result of peak detection with window width = 601

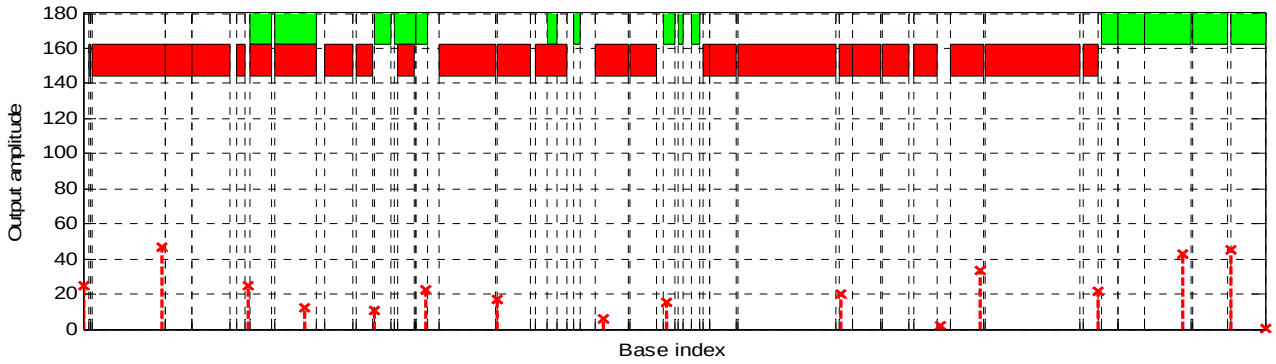

Figure 4. The result of nadir detection with window width = 1601

The length of period-3 sequence $S_i(n), i = 1,2,3,4$ is $L_s = 1950$. The period-3 filter is a sequence with length N = 221. The peak detection window is with length L=601, and nadir detection window is with length L=1601.

The output of period-3 filter is showed in Figure 2. , and the peak detection result is described in Figure 3. The nadir detection result is described in Figure 4.

The regions marked by light green colored tags (first row of tags) is the protein coding region for the 3'-5' sequence, and the dark red tags (second row of tags) is for 5'-3' sequence, the blank regions show non-coding regions for both types of sequence. It can be observed that there are some peaks (dark blue lines with a circle on top) in the non-coding region of the 5'-3' DNA sequence of prokaryotic bacteria E. coli strain MG1655. This means that there are coding regions in the complementary 3'-5' sequence. The nadir (dark red lines with a cross on top) stems in Figure 3 indicate that there must be a non-coding region around for both 5'-3' and 3-5' sequence.

Combining the Figure 3 and Figure 4, an approximate reading frame for this genome segment is obtained in Figure 5.

In Figure 5, the red line with a cross on top indicate non-coding region for both 5'-3' and 3'-5' DNA sequence, while the blue line with a circle on top indicate coding region in either sequence, or both. Therefore, some accurate detection approaches, such as Hidden Markov Models or

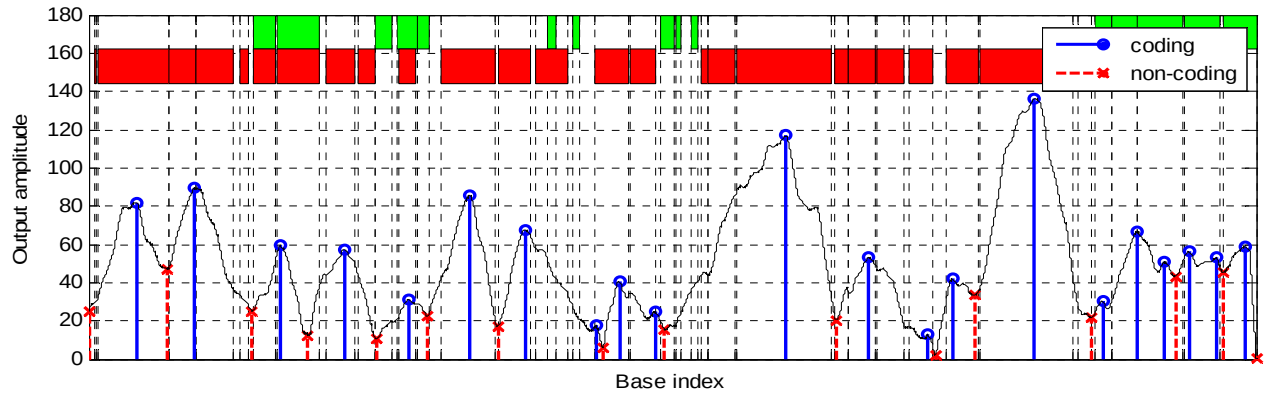

Figure 5. The profile of reading frame.

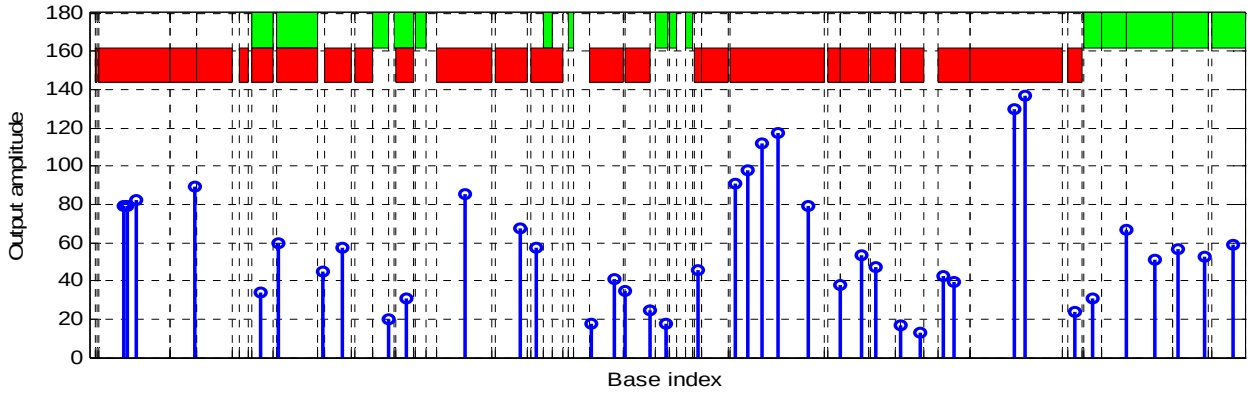

Figure 6. The result of peak detection with narrower detection window.

gene code books, can be used in the areas between neighboring red lines to detect coding region, and the blue lines in these areas mark the locations, where the coding region is most likely located. Obviously, the proposed approach will significantly increase the efficiency in detecting coding regions.

Figure 5. also explained why the 1/3 periodicity has even been observed in non-coding regions for procaryotes. For example, one region is non-coding region for the 5'-3' sequence, but it is coding-region for the 3'-5' sequence. According to [5], the complementary strands are statistically symmetric. Thus, the non-coding region for the 5'-3' will have the same period-3 property as the coding region. It means that the period-3 still works in prokaryotes, and it can be used to detect coding region.

The length of detection window must be determined by trading off the correctness and resolution. Narrower detection window will bring about more peaks thus high resolution, but it will also result in more fake peaks at the same time, because it increases the possibility that the detected peak is only local maximization point instead of real maximization point. This scenario is show below in Figure 6, where the length of peak detection window is L = 201.

It can be found that some peaks appear in non-coding regions. These peaks indicate local maximization points. The length of nadir detection window can be decided in the same way.

IV. Conclusion

In this paper, we have introduced a novel protein coding-region detection algorithm, which is based on period-3 property of coding region. The proposed scheme is more efficient than traditional method, because it waive the need to perform variable length DFT. By analyzing the simulation result, we have shown that, with the help of ORF analysis, gene library and the regulatory sequence set, it is possible to obtain a relatively accurate prediction of the protein coding-regions in genome.

References

[1] E. N. Trifonov and J. L. Sussman, "The pitch of chromatin DNA is reflected in its nucleotide sequence," Proc. of the Nat. Acad. Sci., USA, vol. 77, pp. 3816–3820, 1980.

[2] Xiu-Feng Wan, Dong Xu, Andris Kleinhofs, Jizhong Zhou, "Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes", BMC Evolutionary Biology, vol. 4, no. 19, 2004.

[3] H. Herzel, E. N. Trifonov, O. Weiss, and I. Groβe, "Interpreting correlations in biosequences," Physica A, vol. 249, pp. 449–459, 1998.

[4] W. Li, "The study of correlation structures of DNA sequences: A critical review," Computers Chem., vol. 21, no. 4, pp. 257–271, 1997.

[5] Pierre-François Baisnée, Steve Hampson, Pierre Baldi, "why are complementary DNA strands symmetric", Bioinformatics, vol. 18, no. 8, pp. 1021-1033, 2002.