IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 3, Issue 3 (Sep. – Oct. 2013), PP 44-52
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

Structural Analysis and Identification of Genomic and Proteomic Sequences Using Signal Processing Techniques

G.Santhi Kumari[1] , Dr.P.Srihari[2] , K.S.N.V.Someswara Rao[3],
D.Naresh Kumar[4],M.Sujatha[5],H.Shalini[6]

[1]*(Department of E.C.E, M.tech, DIET College of Engineering, India)*
[2]*(Department of E.C.E, Asst.Professor, NITK Surathkal,India)*
[3]*(Department of E.C.E, Asst.Professor, DIET College, India)*
[4]*(Department of E.C.E, Assoc. Professor, LIET College, India)*
[5]*(Department of E.C.E, Asst. Professor, LIET College, india)*
[6]*(Department of E.C.E, M.tech, DIET College of Engineering, India)*

ABSTRACT: *Bioinformatics is a data rich field which provides unique opportunities to use computational techniques, to understand and to organize information associated with Bio molecules such as DNA, RNA, and Proteins. DNA sequences can be converted into RNA and then proteomic sequences using transcription and translation. First these sequences are to be translated into a "gap sequence" consisting of integer numbers. In this paper, Digital signal processing techniques such as Parametric and Non Parametric methods, filtering techniques, transformation domain methods are applied to gap sequences in order to extract gene features such as identification of protein coding DNA regions, identification of reading frames, and similarity between two genomic and proteomic sequences. Extensive experimental results are presented to demonstrate the performance of the method.*

Keywords: *Auto-regressive, Genomic-signal-processing, matched filter, Period gram, protein coding region.*

I. INTRODUCTION

Genomic sequences are composed of symbols. For example, there are four kinds of deoxyribonucleic acid bases, adenine (a), cytosine (c), guanine (g), and thymine (t), constructing the DNA sequences that bear the heritage information. Protein has 20 amino acids .Same DNA can be converted into protein by using transcription and translation as shown in Fig 1.

In this way, both DNA and protein sequences can be represented by an alphabet set containing a finite number of characters.

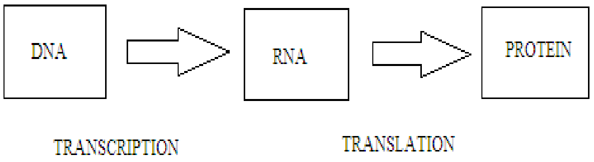

Figure 1: Central Dogma

The genomic sequences to be converted into codons as shown in Fig 2,which contain multiple of three bases. Codon is a Amino Acid, the group of amino acids are protein sequences.

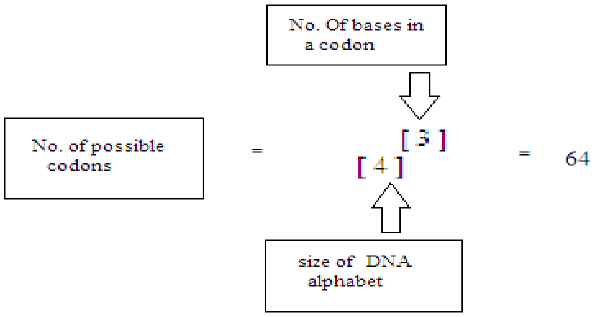

Figure 2: Possible Codons

The set of symbols does not have a well-defined algebraic structure. By converting the symbol sequence into another sequence consisting of numbers, it is possible to obtain more insights into the sequence using the algebraic structure of numbers. One intuitive way to convert symbols into numbers is through a direct fixed mapping. This method is widely used because of its simplicity. Then, they attempted to find the similarity of two symbol sequences by correlating their corresponding complex sequences. The direct mapping method [1] that translated the DNA character set {A,C,G,T} into the integer set {1,2,3,4}. There are other direct mapping methods translated a genomic sequence into a sequence of vectors [8].

In this work, we propose another way to map a symbol sequence to a numerical sequence. We consider the gap sequence that describes the distance between consecutive symbol patterns existing in genomic and protein sequences. The pattern can be a single genomic symbol, a genomic word, or **a** morphological pattern that selects each recognized hit. Through pattern filtering, the genomic sequences are transformed into gap sequences of non-negative integers. The gap sequences abstractly represent the original sequences in the form of gaps, which can be used for the purpose of matching, alignment, and similarity measurement.

Digital Signal Processing (DSP) applications in Bioinformatics have received great attention in recent years, where new effective methods like filtering techniques ,transformation methods , parametric and non-parametric methods are used for genomic sequence analysis[5], such as the detection of coding regions, identification of reading frames have been developed. A matched filtering approach will be presented in Section VI to detect phenomena interested. A matched filter is used to find similarity between AF324494 and AF320294.

Both are converted into protein sequences have similarity located at $I_{MAX=}$588. The match between gap sequences is called a "frame match "or a "structural match". The actual match of two genomic sequences demands both frame match and stuffing match. The proposed approach is useful for sequence analysis based on the frame match with desirable patterns. Obtaining the patterns present in the DNA and protein sequences as well as to develop efficient feature extraction method for achieving better classification.

The obtained results justify the use of the gap sequence as an effective tool for genomic DNA and proteomic sequence analysis. Beyond that, the gap structure can go further to the core issues of the DNA encoding such as morphological DNA structure, sequence decomposition, and advanced pattern filtering. To conclude this work, some discussion is made in Section VI and concluding remarks are given in Section III and V. Extensive experimental results will be presented to demonstrate the performance of the proposed method.

II. GAP SEQUENCES FOR GENE

To translate the sequence in symbols to the sequence in numbers, we should avoid tagging the elements of the mapped sequence with properties that exclusively belong to the numbers.The operations 'greater-than' and 'less-than' are not meaningful to the symbols, but are well defined in various kinds of numbers. Or the other hand, the operations 'equal-to' and 'not-equal-to' work for both the symbols and the numbers. We consider only the meaningful operations during the mapping process to stay away from the problem caused by direct mapping.

A structure mapping technique, pattern filtering, is introduced here which keeps only the structural information in the translated sequence. We start out with an easy case of pattern filtering. Let S be a DNA sequence of length n. S[i] is a DNA character at location i of the sequence *S,* ie, S[i] E{a, c , g , t} **,** i = 1,2,. . . , n.

To point out the locations of some specific character, say 'a', in the sequence *S,* we need an indicator sequence, which is defined as :

$$\bar{I}_a[i] = \begin{cases} 1, \text{ if } s[i] = \text{'a'}, \\ 0, \text{if } s[i] \neq \text{'a'}, \end{cases} \quad \text{where } i=1,.......n\text{-} j_a+1 \;. \quad (1)$$

The pattern length, ***j,,*** is unity in this case. Now we have a binary indicator sequence which is the intermediate step on the translation to sequence of numbers. The pattern filtering is to read the gap between two successive occurrences of some specific character or pattern. To precisely describe [1] the location of the specific pattern we added two virtual values '1' to the head and the tail of $\bar{I}_a[i]$**.** The modified indicator sequence is:

$$I_a[i] = \begin{cases} \bar{I}_a[i] \;, & \text{where } i=1,.......n\text{-} j_a+1 \;. \\ 1, & \text{where } i=0 ....n\text{-} j_a+2 \end{cases} \quad (2)$$

Finally the 'a' filtered sequence $F_a[i]$ is defined as the number of steps from the i^{th} '1' to the $(i+l)^{th}$ '1' in I_a**,** where *i* = 0,1, . n_a**.** All elements in F_a are positive integers, and $n_{a,}$ is the number of ocurrences of the selected pattern 'a'.

The relationships between the sequences above are demonstrated in Fig 3.

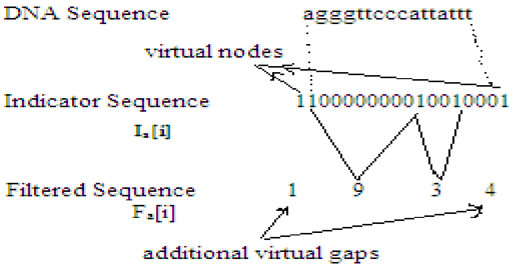

Figure 3: The relation between DNA sequence, the indicator sequence, and the filtered gap sequence

III. Main Dsp Algorithms Employed In The Analysis Of Genomic Sequences

In this section a synthetic overview of the main DSP algorithms that have been used in the analysis of genomic sequences is presented. There are excellent books on DSP theory by Oppenheim and Schafer [2] and Proakis and Manolakis [3].

3.1. Discrete Fourier Transform

The Discrete Fourier Transform is a mathematical operation that transforms one discrete, limited (finite) N duration function into another function, according to:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-i\frac{2\Pi nk}{N}} \quad ,0 \leq n,k \leq N-1 \qquad (3)$$

The function X[k] is the Discrete Fourier Transform (DFT) of the sequence x[n] and constitutes the frequency domain representation of x[n], which is usually (or conventionally considered) a function in the time domain. The Discrete Fourier Transform only evaluates the frequency components required to reconstruct the finite segment of the sequence that was analyzed. In general, the DFT is a function in the complex domain as a result of the complex exponential in the right side of equation (3), and for the particular case of real sequences, it will be a sequence of complex numbers of the same length as x[n].

The DFT, however, suffer from three important drawbacks as a tool for spectral analysis: a) Spectral leakage, which means the presence of energy in zones where the spectrum should be zero [7] b) the frequency response of the DFT coefficients is not constant with frequency ("picket-fence" effect), and c) the spectral resolution, or ability to separate frequency lines that are close in frequency, depends inversely upon the length of the sequence in the time domain. This means that the DFT cannot distinguish appropriately close spectral components for time signals of short duration.

Multiplying the time signals by special weighting functions called windows, and controlling the signal length, can help in overcoming these limitations in some extent. According to [8, 11], a protein coding region exhibits a peak at k = N / 3 (exhibits a relatively large value) whereas no such behavior is observed in the non-coding regions. This property is referred to as "periodicity property "or "period-3" behavior. Here, we implement one of those discrete Fourier transform (DFT) based splicing algorithm mentioned in [5]. Period-3 behavior is exploited in the DFT based splicing algorithm to find the protein coding regions in a DNA sequence. Here, we compute the magnitude of the frequency component at k = N / 3. Finally, we calculate the total strength of the peak (Exons). The total strength of the peak in the protein coding region can be calculated as [10]:

$$S[k]=X_A[k]^2+X_T[k]^2+X_c[k]^2+X_G[k]^2 \qquad (4)$$

The traditional DFT approach loses its electiveness in case of small DNA sequences for which the autoregressive (AR) modeling has been used as an alternative tool.

3.2. Spectral Analysis Using Non Parametric Method

Non-parametric technique of spectrum estimation is based on the idea of first estimating the auto-correlation of data sequence and then taking its Fourier Transform to obtain its Power Spectral Density (PSD).This method also known as Periodogram method was first introduced by Schuster in 1898,in his study of periodicities in sunspot numbers. Although periodogram is easy to compute it is limited in its ability to produce an accurate estimate of the power spectrum, particularly for short data records [4]. For improvement of statistical property of periodogram method a variety of modifications have been proposed such as Bartlett's method, Welch's method and the Blackman-Tukey method. In periodogram method PSD is estimated directly from signals itself [13].The Fourier Transform of the estimated auto-correlation of data sequence is given by the following equation:

$$P_x\left(e^{j\omega}\right) = \sum_{k=-\infty}^{\infty} r_x(k)e^{-j\omega k} \qquad (5)$$

The estimated auto-correlation function:

$$r_x(k) = \frac{1}{N}\sum_{k=-\infty}^{\infty} x(n+k)x^*(n) \qquad (6)$$

Where k=0,1,2.........N-1;with $r_x(k)$ set equal to 0 for $|k| \geq$ N.With values of $r_x(k)$, for k<0 defined using conjugate symmetry as: $r_x(-k)= r_x^*(k)$

Taking DTFT of $r_x(k)$ leads to the estimate of the power spectrum known as Periodogram:

$$P_{per}\left(e^{j\omega}\right) = \sum_{k=-N+1}^{N-1} r_x(k)e^{-j\omega k} \qquad (7)$$

The Periodogram power spectrum estimation of two gap sequences (AF320294 and AF324494) of Fig (3) is shown in Fig 4(a),4(b).

3.3.Spectral Analysis Using Parametric Method

The Parametric method uses a different approach to Spectral estimation. Instead of estimating PSD from data directly as is done in non-parametric method, it models the data as output of a linear system driven by white noise and attempts to estimate parameters of this linear system. The most frequently used linear system model is the all pole model, a filter with all of its zeroes at the origin on the z-plane. The output of such a filter for white noise input is an AR process, known as AR method of spectral estimation. There are different types of AR methods such as Burg method, Covariance and Modified Covariance method, Yule-Walker (auto-correlation) method etc. The advantage of Yule-Walker Autoregressive method is that it always produces a stable model

Parametric methods can yield higher resolution than non-parametric methods when the signal length is short [12], [6]. As already stated the signal spectrum estimated in parametric method is based on the PSD of a linear system driven by white noise.

The output of such a system with white noise input referred to as Autoregressive (AR) process has been implemented here (Fig 5). The pth order power spectrum of Auto Regressive process is given by equation (8):

$$P_{AR}\left(e^{j\omega}\right) = \frac{|b(0)|^2}{|1 + a_p\ (k)e^{-j\omega k}\ |^2} \qquad (8)$$

Where b(0) and ap(k) are estimated from given data

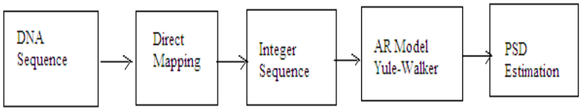

Figure 5: Block diagram realization of an AR model PSD estimation system

Here Yule-Walker Autoregressive method has been implemented efficiently for Parametric Analysis of DNA sequence. AR models are popular because with them an accurate estimation of PSD can be obtained by solving linear equations. Since in above equations $|b(0)|^2$ is constant, the only value that are needed for calculating the shape of PSD are the coefficients $a_p(k)$.

The Yule-Walker (auto-correlation) method has been used here for its simplicity. Yule-Walker power spectrum estimation of two gap sequences (AF320294 and AF324494) of Fig (3) are shown in Fig 6(a),6(b). From that exons, the codons (group of Amino Acids)are generated.

IV . GAP SEQUENCES FOR PROTEIN

The Protein sequences can be developed from DNA sequences.DNA to Codon and then protein.Codon sequences& reverse Codon sequences are shown in Fig 7 (a, b, c).

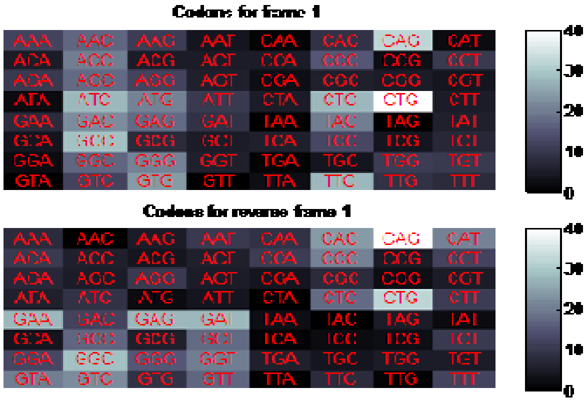

Figure 7(a): Codons For Frame 1 & Reverse Frame1

There are six different ways[8] in DNA to read the sequence. Each one is called a reading frame as shown in Fig 8.

CGT AGC TTA CTG . . .
. CG TAG CTT ACT G. . .
. . C GTA GCT TAC TG.

Figure 8: Three ways of reading DNA strand.

The remaining codons are

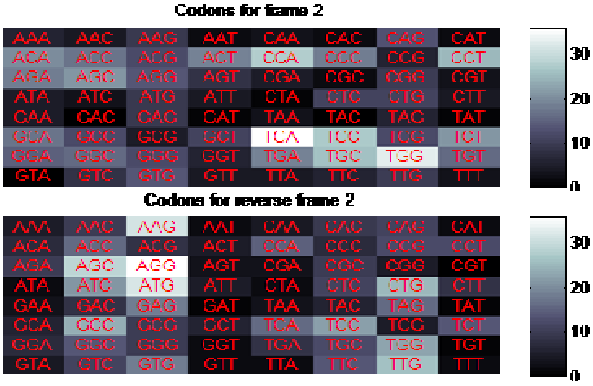

Figure 7(b): Codons For Frame2 & Reverse Frame2

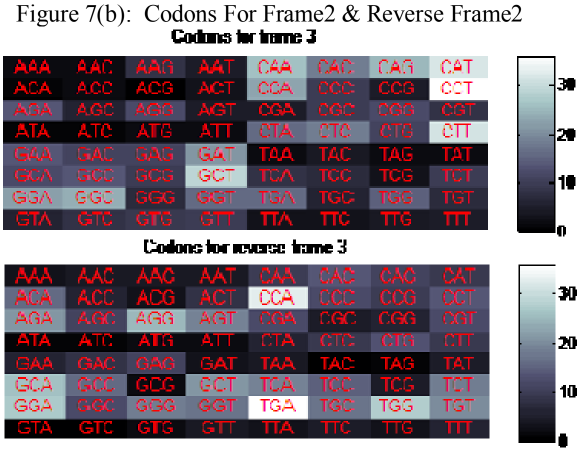

Figure 7(c): Codons For Frame3 & Reverse Frame3

The letters A, T, C, and G represent molecules called Nucleotides or bases .Since DNA contains the genetic information of living organisms, we see that life is governed by quaternary codes. Another example of discrete-alphabet sequences in life forms is the protein. A large number of functions in living organisms are governed by proteins. A protein can be regarded as a sequence of amino acids.The twenty letters used to denote the amino acids are the letters from the English alphabet except B, J, O, U, X, Z.

For example a part of the protein sequence could be

. . . PPVAC ATDEEDAF GGAYPQ . .

If we assign numerical values to the four letters in the DNA sequence, we can perform a number of signal processing operations such as Fourier transformation, digital filtering, power spectrum estimations. Some of those are quite interesting and in fact have important practical applications. Similarly, once we assign numerical values to the twenty amino acids in protein sequences, we can do useful signal processing. Some of Amino Acids with nucleotide code can be represented by

For Example:

Alanine (A)is GCT,GCC,GCA,GCG;

Arginine(R) is CGC, CGA, CGG, AGA, AGG;

Asparagines (N) is ATT, AAC

Similarly the gap sequence of protein can be taken as DNA's gap sequence as shown in Fig 5. Similarly Indicator sequence developed for protein sequences as same as indicator sequence of Genomic sequences which is shown in Fig9.

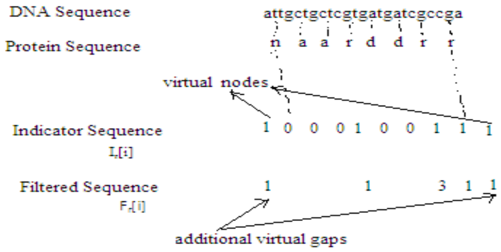

Figure 9: The relation between DNA & Protein sequence, the indicator sequence, and the filtered gap sequence

V. Matched Filter For Genomic And Proteomic Sequences

When we measure the similarity between two sequences, the correlation operation is a good measurement. We are not able to pick up the similarity if we correlate the sequences in the head-to-head manner; i.e., correlating from the beginning of both sequences. To allow checking in every possible starting location, the correlation must be collected in every possible shift between the two sequences. This is equivalent to convoluting these two sequences in the opposite direction[1].

We describe as the following two objectives of matched filter.

1) Is there a significant similarity between the query sequences and target sequence?

2) If yes, where is the similarity location in both sequences?

We can expect to obtain some spike(s), denoted by C[i], in the matched filter output by the convolution operation for the query and the target sequences. To check the significant similarity, a detection threshold will be set to sweep out insignificant similarities. Regarding to finding the locations of similarities, the location of the spikes will indicate the amount of corresponding shifts between the query and the target sequences. At the matched filter output is a collection of different shifting-amount correlations. Let $F_t[i]$ and $F_q[i]$ be the target and the query gap sequences, respectively. The matched filter output is:

$$C[i] = Ft[i] \otimes Fq[i] = \sum_{k=-\infty}^{\infty} Ft[i]Fq[i+k] \qquad (9)$$

n_q and n_t denote the length of corresponding gap sequences $F_q[i]$ and $F_t[i]$, the length of C[i] could be as long as $n_t + n_q - 1$.

The time complexity to compute C[i] is $O(n_t+n_q)$, including the FFT, multiplication in Fourier domain, and IFFT. The result is shown in Fig. 6(a), from which we can find at least two problems. The first one **is** that the raw output signal sequence of matched filter is not normalized, which makes the decision of spike locations more complicated. Another problem is the edge effect at both beginning and end portion of the output signal due to the insufficient correlating points. The edge effect introduces high variation to the C [i] values close to the head and tail of the whole signal.

Two problems have been pointed out by judging the result signal sequence at the output of matched filter in Fig. 6(a).One is that the output signal is not normalized, and the other is the edge effect appearing at the head and the tail portion of the output signal.

Two processes are designated to solve these problems. Before fed info the matched filter, we first apply the normalization process to both query gap sequence F_q and target gap sequence F_t, Then use the proposed edge effect reduction process with the output signal of matched filter to reduce the edge effect.

5.1.Correlation Enhancement and Convolution Method

To the normalization, what should be done is to remove the bias component from both $F_t[i]$ and $F_q[i]$ denoted by m_t and m_q, respectively. The bias component of a sequence is the sequence mean. We have the matched filter collecting the cross-covariance rather than the cross-correlation, as in Fig. 6(b), by setting

$C[i]=(F_t[i]-m_t)\otimes(F_q[-i]-m_q)$.

The result of the edge effect reduction process is presented Fig. 6(c) shows convolving one signal sequence with opposite direction of another signal sequence[1] and Fig. 6(d) shows the resulting C'[i] of the correlation enhanced matched filter, which applies the combination of both normalization and edge effect reduction processes.

When same DNA sequences such as AF320294 and AF324494 can be converted into PROTEIN Sequences have same matched filter characteristics used to match ,the spike locates at I_{max}=588 and to predict the sequence analysis which as shown in Fig(10).

VI. Results & Discussion

The nonparametric Power Spectral estimation method is methodologically straight forward and computationally simple. But in case of low Signal to Noise Ratio (SNR) spectral features are difficult to be distinguished and noise artifacts appear in spectral estimates. The traditional DFT approach loses its effectiveness in case of small DNA sequences.

Parametric spectrum estimate methods have more statistical consistency even on short data segments.The comparision of Parametric and Non-Parametric methods for two AF320294 and AF324494 Homo Sapiens genes, Simulation results are

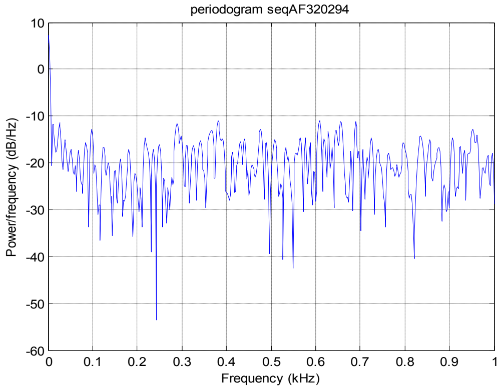

Figure 4(a). Periodogram Power Spectrum Plot For exon of accession no.AF320294

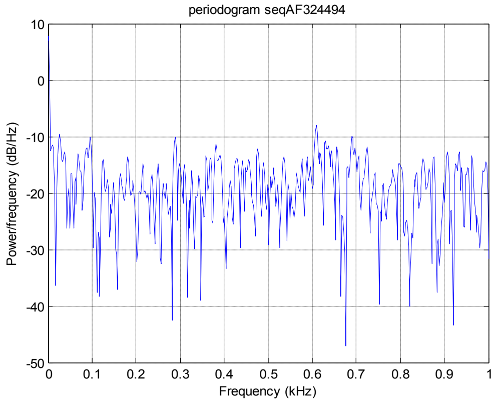

Figure 4(b). Periodogram Power Spectrum Plot For exon of accession no.AF324494

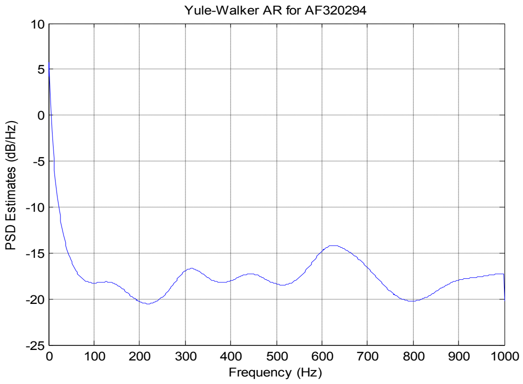

Figure 6(a). Yule-Walker Power Spectrum Plot For exon of accession no.AF320294

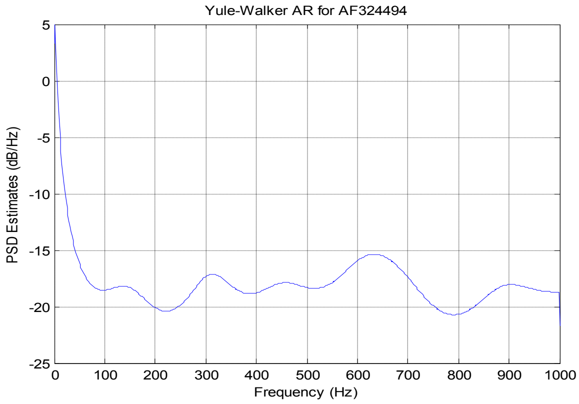

Figure 6(b). Yule-Walker Power Spectrum Plot For exon of accession no.AF324494

The matched filtering approach to find similar segments between gap sequences of Genomic & Proteomic sequences . Simulation results are

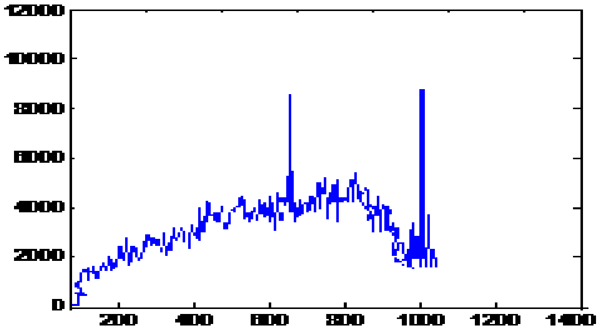

Figure10(a) is the original output signal from matching AF320294 with AF324494. C[i] = *L* =424 + 609 - 1 = 1032. The spike locates at *i,,,* = 588.

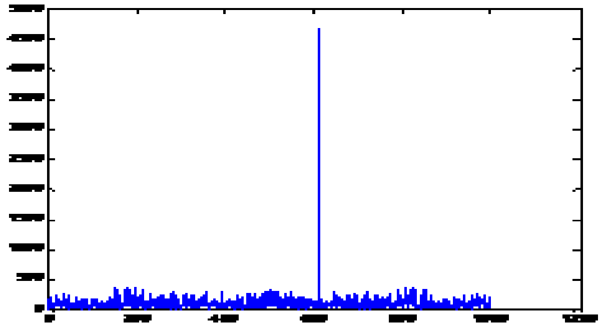

Figure10 (b) the original output has been enhanced by reducing edge effect,

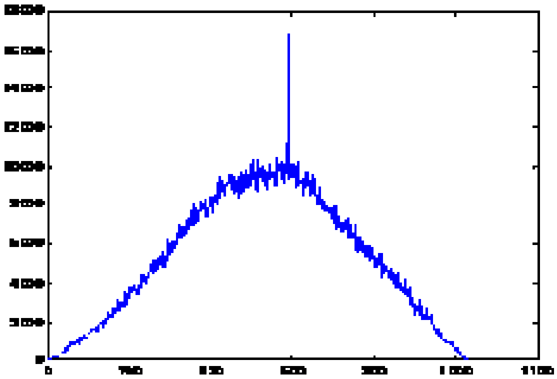

Figure 10(c) is the normalized version of original output.

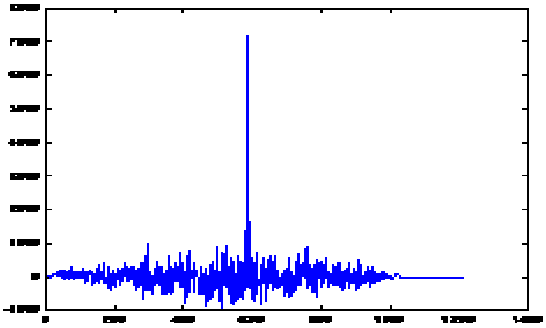

Figure10(d) has been through both of these processes.

Figure 10: The matched filter Output for gene and Protein sequence output and processing thereafter.

Both sequences can be downloaded from Gen Bank at National Center for Biotechnology Information NCBI, http://www.ncbi.nlm.nih.gov/)

VII. CONCLUSION &FUTURE WORK

In this paper ,parametric as well as non-parametric Power spectrum estimation and DFT techniques to find coding region of DNA sequence taken from Homo Sapiens genes.AR model Yule Walker method presented here are popular because they provide accurate estimation of PSD by solving linear equations compared to Periodogram and DFT techniques. After finding exons, used to covert DNA to proteomic sequences. To find similar segments between gap sequences of DNA and proteomic sequences matched filtering approach is applied. By detecting spikes in the filtered output, we are able to locate and align similar segments between two sequences. Future course Genes from other species may also be taken into consideration. Simulation results demonstrated the good performance of the proposed scheme, including accurate results and a fast processing speed. In the near future, we would like to extend the technique to find some desirable patterns in genomic sequences, Spectral analysis of various signals can provide useful material for diagnosis.

REFERENCES

[1] *Shih-Chieh Su, Chia H. Yeh and C.-C Jay Kuo ,Structural Analysis of Genomic Sequences with Matched Filterin,,*2003:17-21

[2] Oppenheim AV, Schafer R. *Discrete-Time Signal Processing (3rd Edition)(*Prentice-Hall, NY 2009.)

[3] Proakis J G, Manolakis DK, *Digital Signal Processing (4th Edition)(* Prentice Hall, NY 2006).

[4] Stoica P, Moses RL, *Spectral Analysis of Signals,* Prentice-Hall, NY 2005.

[5] Akhtar M, Epps J, Ambikairajah E, *Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction, IEEE J Select Topics Sign Proc* 2008; 3: 310-21.

[6] Hayes M.H., *Statistical digital signal processing and modeling,* John Wiley & Sons, Inc., New York, USA, 1996.

[7] Juan V. Lorenzo-Ginori , *Digital Signal Processing in the Analysis of Genomic Sequences* 2009:28-40

[8] Anastassiou, D,*Genomic Signal Processing,* IEEE Signal Processing Magazine 18,no. 4 2001):8-20.

[9] E. A. Cheever, D. B. Searls, W. Karunaratne*, and G. C. Overton,Using signal processing techniques for DNA sequence comparison, in Bioengineering Conference, 1989, pp. 173–174.*

[10] Vaidyanathan, P. P. and Byung-Jun Yoon.,*Gene and Exon Prediction Using Allpass-Based Filters.,* In IEEE International Workshop on Genomic Signal Processing and Statistics, CP2-02. Piscataway, NJ: IEEE Press, 2002

[11] Tiwari, S., S. Ramachandran, A. Bhattacharya., S. thattacharya, and R. Ramaswamy.,*Prediction of Probable Genes by Fourier Analysis of Genomic Sequences*, Computer Applications in the Biosciences 113, no. 3 (1997):263-270.

[12] Chakrabarty Niranjan, Spanias A., Lesmidis L.D. and Tsakalis K., *Autoregressive Modeling and Feature Analysis of DNA Sequences,* EURASIP Journalon Applied Signal Processing 2004:I, 13-28.

[13] Nair Achuthsankar. S. and Mahalaxmi T.*, Are Categorical periodograms and Indicator sequences of genomes spectrally equivalent?,* In silico Biology 6, 0019 (2006) Bioinformatic Systems', v.