OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Identifying the Coding and Non Coding Regions of DNA Using Spectral Analysis

R. Vijayashree[1], D. Naresh Kumar[2], P. Madhuri[3], S. K. Asha Begum[4], P. Sumanth[5]

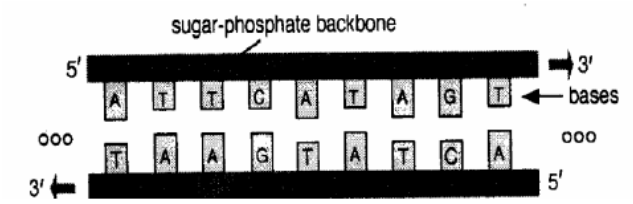
[1, 2, 3, 4, 5] *Dept. of ECE, Lendi Institute of Engineering and Technology, Affiliated to JNTUK, Vizianagaram*

Abstract: *This paper presents a new method for exon detection in DNA sequences based on multi-scale parametric spectral analysis Identification and analysis of hidden features of coding and non-coding regions of DNA sequence is a challenging problem in the area of genomics. The objective of this paper is to estimate and compare spectral content of coding and non-coding segments of DNA sequence both by Parametric and Non-parametric methods. In this context protein coding region (exon) identification in the DNA sequence has been attaining a great interest in few decades. These coding regions can be identified by exploiting the period-3 property present in it. The discrete Fourier transform has been commonly used as a spectral estimation technique to extract the period-3 patterns present in DNA sequence. Consequently an attempt has been made so that some hidden internal properties of the DNA sequence can be brought into light in order to identify coding regions from non-coding ones. In this approach the DNA sequence from various Homo Sapiens genes have been identified for sample test and assigned numerical values based on weak-strong hydrogen bonding (WSHB) before application of digital signal analysis techniques.*

I. Introduction

The enormous amount of genomic and proteomic data that are available in public domain inspires scientists to process this information for the benefit of the mankind. The genomic information is present in the strands of DNA and represented by nucleotide symbols (A, T, C and G).The segments of DNA molecule called gene is responsible for protein synthesis and contains code for protein in exon regions within it. When a particular instruction becomes active in a cell ,the corresponding gene is turned on and the DNA is converted to RNA and then to protein by slicing up to exons (protein coding regions of gene).Therefore finding coding regions in a DNA strand involves searching of many nucleotides which constitute the DNA strand. As the DNA molecule contains millions of nucleotide element, the problem of finding the exons in it is really a challenging task. It is a fact that the base sequences in the protein coding regions of DNA molecules exhibit a period-3 pattern because of the non uniform distribution of the codons in recent past many traditional as well as modern signal processing techniques have been applied to process and analyze these data have used DFT for the coding region prediction.

The DFT based spectrum estimation methods produce the windowing or data truncation artifacts when applied to a short data segment. The Parametric spectral estimation methods, such as the autoregressive model, overcome this problem and can be used to obtain a high-resolution spectrum. But rapidly acquiring the genomic data demands accurate and fast tools to analyze the genomic sequences. In this paper we propose an alternate but efficient and cost effective technique for the identification of the protein coding regions exhibiting period-3 behaviour. These new methods employ the adaptive AR modelling which require substantially less computation and yield comparable performance than the conventional approaches



DNA double helical structure

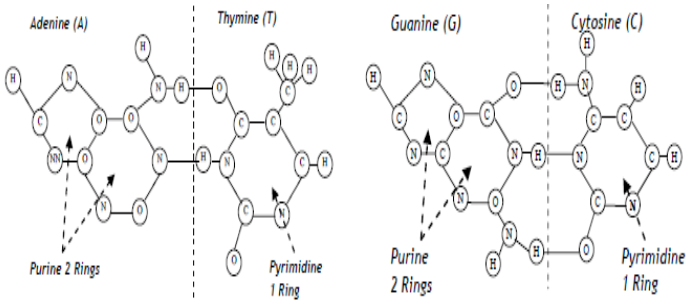
Only around 1.2 % of human DNA is known to be coding for proteins. Our knowledge of the role and location of other elements is limited and new types of sequences unknown function are still discovered. Recently, several sets of highly conserved non coding sequences have been identified in vertebrate genomes. A combination of comparative genomic studies and laboratory experiments has shown that these conserved non-coding elements (CNEs), most of which are more conserved than protein-coding exons.

Most of these sequences are located in and around developmental regulation genes and when some of them were tested in the laboratory, they appeared to drive tissue-specific gene expression in early development.

Genome sequences contain the genetic information of living organisms. This information, which is common to all life, is coded in the deoxyribonucleic acid (DNA) sequences. Understanding the codes will enable us to understand the life processes of a particular organism. As such, even with the genome sequence in hand, much work remains to be done to lay open the genetic secrets of a particular species.

II. DSP Techniques For Spectral Estimation Of DNA Sequences

Decoding the meaning of the nucleotide characters A, T, C, and G is becoming even more pressing with the final release of the sequencing of the human genome.



Double hydrogen bond Triple Hydrogen bond A=T signifies weak bond c=g signifies strong bond

Gene identification is of great importance in the study of genomes, to determine which open reading frames (ORFs) in a given sequence are coding sequences for prokaryotic, and to determine the exons and introns, and the boundaries between them in a given gene for eukaryotic DNA sequences. There are a number of identification methods being used, either with training datasets, or without any database information. Gene scan use a semi-hidden Markov model, and FEX use a linear discriminant function to determine genes, are examples of gene or exon finding algorithms based on database information.

Examples of algorithms without database information are statistical correlation analysis statistical regularity to detect coding regions and Fourier analysis.

For example a DNA sequence of length N:

X [n] = [A T G C C T T A G G A T](1)

After mapping:

Xsw [n] = [2 2 3 3 3 2 2 2 3 3 2 2]

Among the various methods, the most prominent distinctive feature of coding and non-coding regions is the 3 base pairs (bp) periodicity or 1/3 frequency,which has been shown to be present in coding sequences? The periodicity is caused by the coding biases in the translation of codons into amino acids.

The Fourier transform analysis has been widely used for sequence processing However, Fourier transform contains the problems of windowing or data truncation artifacts and spurious spectral peaks, and thus, the spectral obtained using the Fourier transform will exhibit the same problems. This problem has been studied extensively in digital signal and image processing, where autoregressive (AR) models are used to achieve a high spectral resolution. The AR model or linear prediction (LP) process is a relatively new approach to spectral analysis to overcome the limitation of Fourier methods.

III. Non-Parametric Analysis of DNA Sequence

To perform the gene prediction based on period-3 property the total DNA sequence is first converted into four indicator sequences, one for each base. The DNA sequence D(n) is mapped into binary signals uA(n), uC(n), uG(n) and uT(n)

Which indicate the presence or absence of these nucleotides at location n.

For example the binary signal uA (n), attributed to nucleotide A takes a value of 1 at n = n0 if
D(n0) = A, else uA (n0) is 0. Suppose the DNA sequence is represented as
D(n) = [AT GAT CGCAT]
Then its numerical representation is given by
uA (n) = [1001000010]
uC (n) = [0000010100]
uG (n) = [0010001000]
uT (n) = [0100100001]
Thus uA (n) + uC (n) + uG (n) + uT (n) = 1

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. 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. For improvement of statistical property of periodogram method a variety of modifications have been proposed such as Barlett's method, Welch's method and the Blackman-Tukey method. In periodogram method PSD is estimated directly from signal itself.

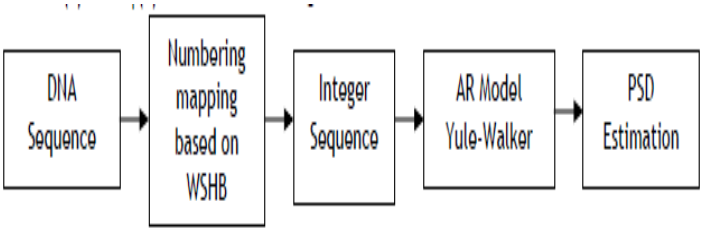
IV. Parametric Analysis of DNA Sequence

Parametric analysis can yield higher resolutions than nonparametric analysis in cases when the signal length is short. These methods use a different approach to spectral estimation; instead of trying to estimate the PSD directly from the data, they model the data as the output of a linear system driven by white noise, and then attempt to estimate the parameters of that linear system.
The most commonly used linear system model is the all-pole model, a filter with all of its zeroes at the origin in the *z*-plane. The output of such a filter for white noise input is an autoregressive (AR) process. For this reason, these methods are sometimes referred to as AR methods of spectral estimation.

The AR methods tend to adequately describe spectra of data that is "peaky," that is, data whose PSD is large at certain frequencies. The data in many practical applications (such as speech) tends to have "peaky spectra" so that AR models are often useful. In addition, the AR models lead to a system of linear equations which is relatively simple to solve. 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.

$$P_{AR}(e^{jw}) = \frac{|b(0)|^2}{|1+a_p(k)e^{-jwk}|^2}$$

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.



Block diagram realization of an AR model PSD estimation system

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 For achieving on line prediction of gene and exon the computational time needs to be reduced. Further the fixed AR method requires all data to be available simultaneously which is not always feasible. With a motive to alleviate these limitations an adaptive AR model based approach is suggested in this section for efficient prediction. The AR process can be viewed as an adaptive prediction error filter. which uses two bases out of the four DNA nucleotides by ignoring the base order, there are six combinations: AC, AC, AG, TC, TG, and CG. The power spectrum as constructed using can be given by $|P(f)|^2$.

4.1 BURG METHOD:

Another type of parametric method is Burg method The Burg method for AR spectral estimation is based on minimizing the forward and backward prediction errors while satisfying the Levinson-Durbin recursion In contrast to other AR estimation methods, the Burg method avoids calculating the autocorrelation function, and instead estimates the reflection coefficients directly.

The primary advantages of the Burg method are resolving closely spaced sinusoids in signals with low noise levels, and estimating short data records, in which case the AR power spectral density estimates are very close to the true values. It is always a stable model

In addition, the Burg method ensures a stable AR model and is computationally efficient.

The accuracy of the Burg method is lower for high-order models, long data records, and high signal-to-noise ratios which can cause **line splitting**, or the generation of extraneous peaks in the spectrum estimate. The spectral density estimate computed by the Burg method is also susceptible to frequency shifts (relative to the true frequency) resulting from the initial phase of noisy sinusoidal signals. This effect is magnified when analyzing short data sequences.

Main characteristics of burg method is Does not apply window to data, Minimizes the forward and backward prediction errors in the least squares sense, with the AR coefficients constrained to satisfy the L-D recursion.

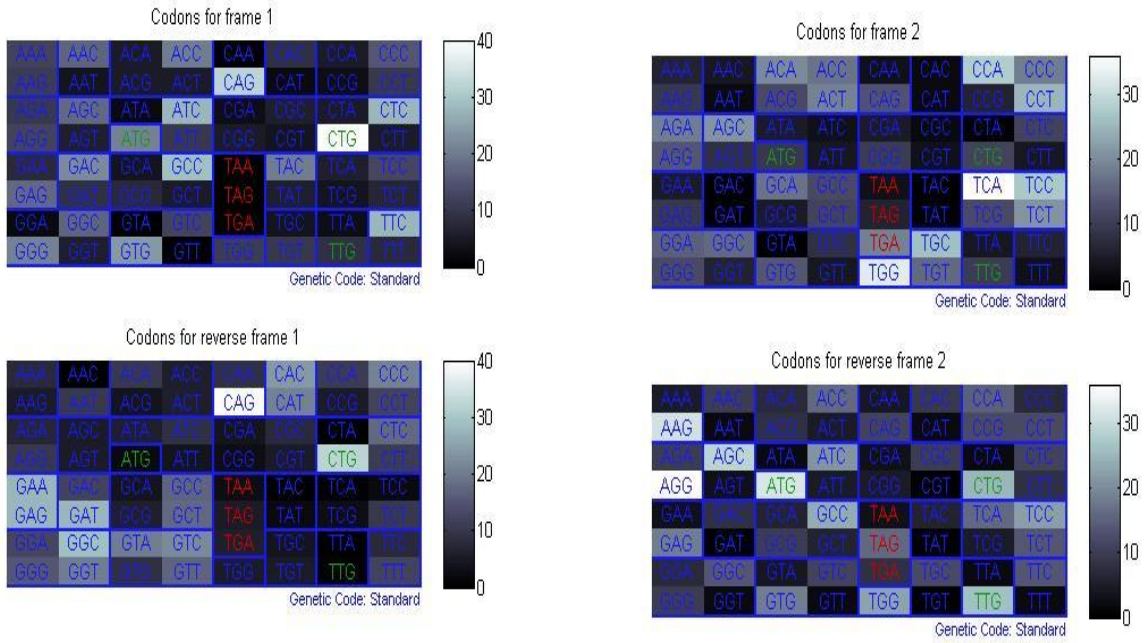
4.2 Covariance and Modified Covarience Methods:

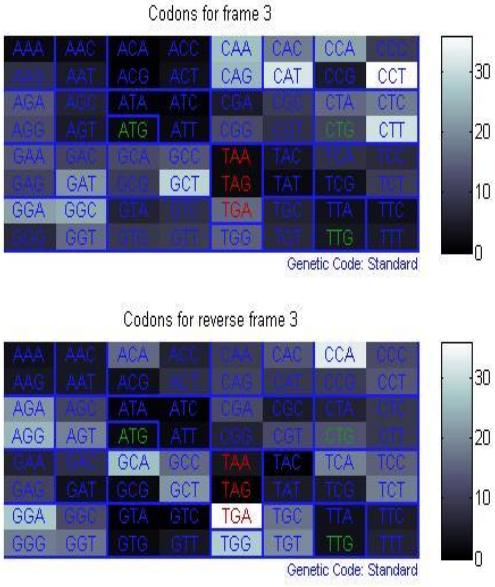
The covariance method for AR spectral estimation is based on minimizing the forward prediction error. The modified covariance method is based on minimizing the forward and backward prediction errors. Does not apply window to data. Major advantages of these methods are better resolution than Y-W for short data records (more accurate estimates). Able to extract frequencies from data consisting of p or more pure sinusoids.

V. Genomic Sequence for Protein

The protein sequences can be developed from DNA to codon and then protein codon sequences & reverse codon sequences are shown in fig (a,b,c).

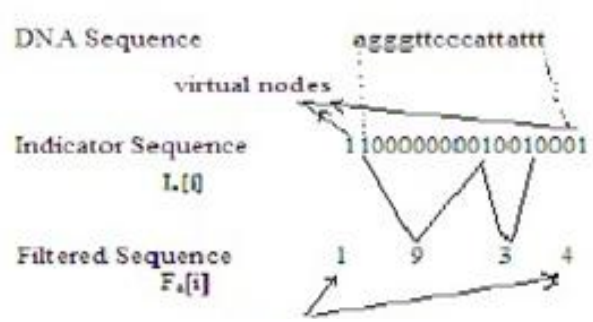
If we assign numerical values to the four letters in the DNA sequence, we can perform a number of signal processing operations such as Fourier transformation, digital filtering, power spectrum estimations.





Some of the 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
Part of a protein sequence could be
...PPVACATDEEDAF GGAYPO..

Similarly the gap sequence can be taken as DNA gap sequence...similarly indicator sequence developed for protein sequences as same as indicator sequence of Genomic sequences...



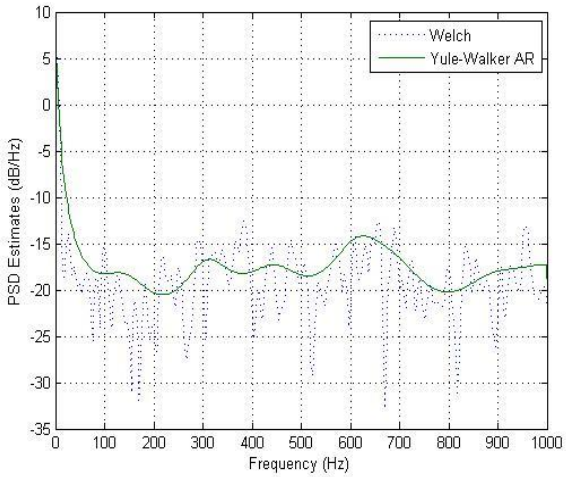
The relation between DNA sequence, the indicator sequence, and the filtered gap sequence

A short analysis frame may detect short exons and introns, but causes more statistical fluctuations. A larger window size may miss the short exons and introns, but cause fewer false negatives and false positives.
Thus, we make use of multiple window sizes, with the aim of reinforcing the advantages of both short and long window sizes but overcome the disadvantages that are caused by them We select the window size within the range of 60bp-360bp The P_{ratio} combination of the windows.

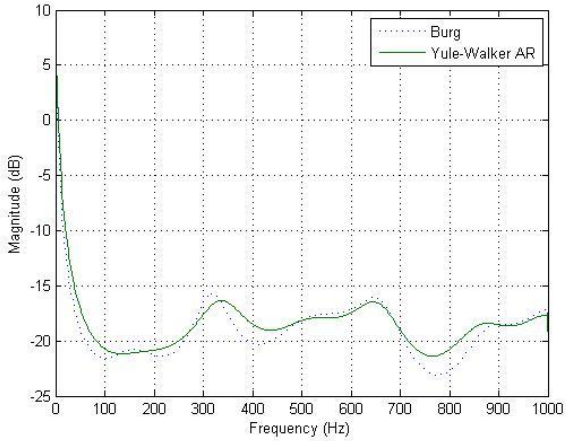
VI. Results and Discussions

6.1 Graphical Representation of Parametric Methods:

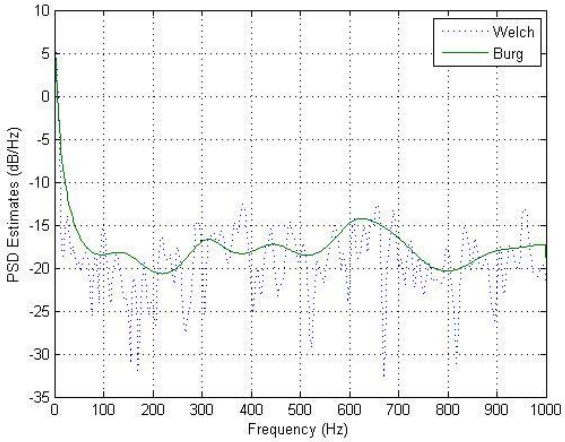
(1) Welch –Yule walker AR



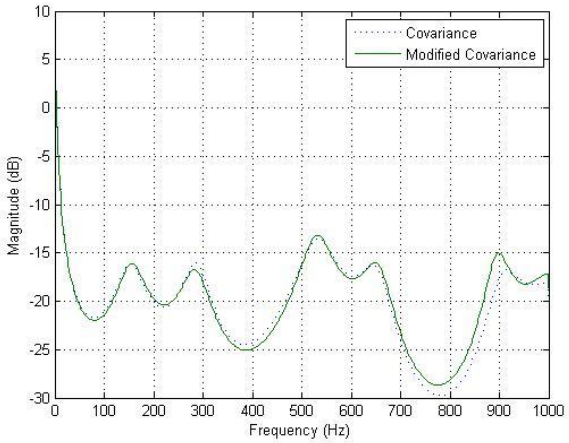
(2) Burg-Yule walker AR



(3) Welch-Burg



(4) Covariance-modified covariance



VII. Conclusion

We are well aware that stochastic or random signals have information bearing Power Spectral Density. Parameterization of a stochastic signal for efficient representation of this information is already in use for speech coding and various other biomedical signal processing applications. In this Paper we have applied parametric as well as non-parametric power spectrum estimation techniques to coding and non-coding regions of DNA sequence taken from Homo Sapiens genes. A comparative study has been organized in order to distinguish coding from non-coding regions. The non-parametric Power Spectral estimation method is methodologically straight forward and computationally simple among several parametric models available, AR models presented here are popular because they provide accurate estimation of PSD by solving linear equations. Data sets from various Homo Sapiens genes have been investigated.

It has been observed from the plots of average PSD for low order Auto Regressive Yule Walker method that the spectral signatures of exons bear a significant pattern as compared to that of introns after finding the exons used to convert DNA to protenic sequences. To find similar segments between gap sequences of DNA.

The normalized variance values show strong periodicities in case of exons for parametric and non-parametric methods where as introns do not reveal any such property. Future course of investigation may be steered towards other parametric methods such as Burg, Covariance, and Maximum Entropy etc.
Genes from other species may also be taken into consideration.

REFERENCES

[1] Anastassiou D., "Frequency–domain analysis of biomolecular sequences", Bioinformatics 16, 1073-1081.
[2] Anastassiou D., "DSP in genomics: Processing and frequency domain analysis of character strings," IEEE, 0-7803-7041-2001.
[3] Chakrabarty Niranjan, Spanias A., Lesmidis L.D. and Tsakalis K., "Autoregressive Modelling and Feature Analysis of DNA Sequences", EURASIP Journal on Applied Signal Processing 2004:I, 13-28.
[4] Ficket J.W. and Tung C.S., "Recognition of protein coding regions in DNA sequences", Nucleic Acids Research, Vol.10, No.17, pp.5303-5318, July 1982.
[5] Ficket J.W. and Tung C.S., "Assessment of protein coding regions in DNA sequences", Nucleic Acid Res, 10, 5303-5318, 2000.
[6] Hayes M.H., "Statistical digital signal processing and modelling", John Wiley & Sons, Inc., New York, USA, 1996.