STM JOURNALS
Scientific Technical Medical

Journal of Artificial Intelligence Research & Advances

Volume 1, Issue 1
www.stmjournals.com

Messenger RNA Approach for Identifying Inter-nucleotide Signals using Genomic Sequences

Pulugurta Krishna Subba Rao*[1], M.S.N. Murthy[2]
[1]Professor, Department of CSE, G.V.P College of Engineering, Madhurawada, Visakhapatnam
[2]Assistant Professor, Department of CSE, G.V.P College of Engineering, Madhurawada, Visakhapatnam

Abstract

The inter-nucleotide signals are a novel way of genomic signal representation of genomic data which is seen to have a discriminatory capability in highlighting the promoter region of gene sequences. The results of applying DFT to in genomic sequence signals that they can discriminate using messenger RNA for identifying inter nucleotide sequences. Genomic Signal Processing (GSP) applications in bioinformatics research have received great attention in recent years, where new effective methods for genomic sequence analysis, such as the detection of converting the genomic sequences into signals have been developed. The use of GSP principles to analyze genomic sequences that requires defining an adequate representation of the nucleotide by numerical values, converting the nucleotide sequences into nucleotide signals. In this paper we present an approach of GSP algorithm which is used in conversion of genomic sequences into signals and also calculating the peaks at coding regions this article also shows the comparative study at the different peaks at a certain elapsed time. The position of these peaks helps in identifying the motifs and the functional behavior of a gene for identifying the Bio Marker selection using a discrete Fourier transform approach

Keywords: *Genomic signal processing, standard genomic sequences method, binary standard selection method, nucleotide sequences, biomarker selection*

****Author for Correspondence*** E-mail: krishnasubbarao@gvpce.ac.in, krishna.pulugurta@gmail.com

INTRODUCTION

This paper is organized in the following way. Firstly an overview of the main DSP algorithms used in applications to genomic sequence analysis is shown: the Discrete Fourier Transform (DFT), the entropy measures, genomic sequences into signals and it also compares the sequence results obtained from the Standard Genomic Signal Processing technique and the enhanced method called Binary Standard Selection Method [1].

The second objective is to identify inter nucleotide signals That is responsible for Type II Diabetes Mellitus Implementing & Analysis of Genomic Signal Process using messenger RNA and proteins that can easily be applied to DNA sequences as they can be viewed as character sequences with a definite alphabet consisting of four letters A, G, C and T. For that, the sequence should be digitalized first, by mapping the characters to suitable numerical values.

The Fourier spectrum analysis and other digital techniques can be applied to the sequence to extract the information contained in it. Genomicl signal processing provides a powerful coding measure, which can differentiate coding regions (exons) from non-coding ones (introns). The methods mainly depend on the peaks obtained through digital analysis, of the coding segments (where N is the length of the segment), which are absent in non coding regions. The method is inherently model independent and doesn't need any training data [2].

Biomarker

A Biomarker is a substance used as an indicator of a biologic state. It is a characteristic that is objectively measured and evaluated as an indicator of normal biologic

process, pathogenic process, or pharmacologic responses to a therapeutic interventation.

As shown in the figure 1. Identification of Biomarker starts from the biological samples collections. As it is information age these biological samples will be in the state of bits and bytes of electronic records. Input to this stat can be achieved either from Proteomics and Genomic analyzed data is kept in large repositories for public use in the interest of research development. In the identification of Biomarker, Bioinformatics plays very major role at the final phase. In this phase different types of analysis can be done with the help of different Bioinformatics analysis techniques. With the help of existing analysis techniques new Biomarkers is a long – term endeavor that proceeds from the basic research to pilot human studies to full scale epidemiological investigations. The past decade has seen extensive research investigation of biomarkers and the beginnings of their practical applications for risk assessment and environment health management. Bioinformatics is the symbiotic relationship between computational and biological sciences. A biomarker may be used to see the response of the treatment for a disease. New biomarkers are being developed to identify individuals at risk for cancer, detect disease earlier, determine prognosis, detect recurrence, predict response to particular agents, and monitor response to treatment (Karim Bensalah, Francesco Montorsi, Shahrokh F. Shariat, 2007) [2]. Biomarkers of disease play an important role in medicine and have begun to play a greater role in drug discovery and development (Aziz, K.J.,2008) [3]. Biomarker helps the detection a disease at early stages of its stages and also helpful in knowing the state of treatment and how body is acting or responding to the medication. Biomarkers are characteristic biological properties that can be detected and measured in parts of the body like the blood or tissue [4–8]. They may indicate either normal or diseased processes in the body. Biomarkers can be specific cells, molecules, or genes, gene products, enzymes, or hormones. Complex organ functions or general characteristic changes in biological structure can also serve as biomarkers. Although the term biomarker is relatively new, biomarkers have been used in pre-clinical research and clinical diagnosis for a considerable time. For example, body temperature is a well-known biomarker for fever. Blood pressure is used to determine the risk of stroke. It is also widely known that cholesterol values are a biomarker and risk indicator for coronary and vascular disease, and that C-reactive protein (CRP) is a marker for inflammation. Genes encoded in DNA are first transcribed into pre-messenger RNA (mRNA) known as primary transcript. Then pre-mRNA is processed to mature mRNA using various forms of modifications of post transcriptional modifications. Then mature mRNA is used as a template for protein synthesis which is known as translation onto a ribosome. (Dobson CM. 2000) [9] Then three neucleotides set is read at a time by matching each codon to its base pairing anticodon to form transfer RNA(tRNA). Then tRNA recognizes the amino acid corresponding to the codon. The sequence thus obtained is protein sequence.

Algorithms Employed in the Analysis of Nucleotide Signals

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

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-i\frac{2\pi}{N}nk} \quad , 0 \leq n,\ k \leq N\text{-}1$$

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 reconstructing 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, 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 is usually represented in terms of the corresponding magnitude and phase functions that constitute the frequency spectrum of the sequence x[n]. The Discrete Fourier transform is a very useful tool, because it can reveal periodicities in the input data. Entropy

Measures are another example of a signal processing concept that has been used in genomic sequence analysis. The concept of entropy is used in signal analysis as a measure of randomness. The first definition of the entropy of a discrete information source (producing a discrete sequence) was introduced by Shannon [6].

$$H(X) = -\sum_{i=1}^{N} p_i \log p_i$$

where pi are the probabilities of the set of values that can take the sequence X, {x1, x2, .. ,xn}.

Numerical Representation of Genomic Sequences

The first approach to convert genomic information in numerical sequences was given by Voss [8] with the definition of the indicator sequences, defined as binary sequences for each base, where 1 at position k indicates the presence of the base at that position, and 0 its absence.

DNA symbolic-to-numeric representations are presented and compared with existing techniques in terms of relative accuracy for the gene and exon prediction problem. Novel signal processing-based gene and exon prediction methods are then evaluated together with existing approaches at a nucleotide level. DSP techniques are applied to DNA sequences following conversion into numerical signals: DNA numerical representation and Feature extraction: In recent years, a number of schemes have been introduced to map DNA nucleotides into numerical values. Some possible desirable properties of a DNA numerical representation include:

- Each nucleotide has equal "weight" (e.g., magnitude), since there is no biological evidence to suggest that one is more "important" than another.
- Distances between all pairs of nucleotides should be equal, since there is no biological evidence to suggest that any pair is "closer" than another. Representations should be compact; in particular, redundancy should be minimized.
- Representations should allow access to a range of mathematical analysis tools.

Applications of GSP

Genomic Signal Processing applications to Bioinformatics started in recent years in which great attention was put to the problem of genomic sequence analysis. Genomic signal processing focuses on the biological mechanisms driving the development of genomic signal processing, in addition to their manifestation in gene-expression-based classification and genetic network modeling.

METHODS USED

Normal Genomic Processing

Method to convert the DNA sequence into a signal: In this method the genes are converted into a from a signal processing perspective, genomic and proteomic data can be viewed as noisy (continuous or discrete) signals that convey valuable information about molecular structure and activities in cells.

- Sequence Analysis: Converting the nucleotide character strings into numerical sequences, computing their discrete Fourier transform (DFT) to determine their unique characteristic frequencies, and changing the amplitudes of the characteristic frequencies in the Fourier spectra to cause a corresponding variation in the time-domain samples of the numerical protein sequence. It is a tetrahedron representation. There are three main dichotomies of the nitrogenous bases biochemical properties that allow arranging them in classes:
  - Molecular Structure—A and G are purines (R), while C and T are pyrimidines (Y);
  - Strength of Links—bases A and T are linked by two hydrogen bonds (W—weak bond), while C and G are liked by three hydrogen bonds (S—strong bond); (3)
  - Radical Content—A and C contain the amino (NH_3) group (M class), while T and G contain the keto (C=O) group (K class).

Binary Standard Selection

Method to convert the DNA sequence into a signal: This method is used reviewed the molecular basis of cancer and our current understanding on genomics and proteomics of cancer. DNA and protein compliments of cancers that dictate the subsequent disease phenotype would eventually lead to breakthroughs. The impact of modern

technology on cancer diagnosis, prognosis, and treatment will also be discussed.

- Sequence Analysis: Many different design formats of gene expression assays include SAGE, complementary DNA (cDNA) arrays, fiber optic arrays (e.g., Illumina), short oligonucleotide array (e.g., Agilent inkjet), and long oligonucleotide arrays (e.g., Asymetrix).
- Normalization reduces variation in signal intensity between spectra. A commonly used normalization method for mass spectrometric data is rescaling each spectrum by its total ion current, i.e. the area under the curve.

The relative Study: The outputs table representations of the time an elapsed the standard Method over the new Binary Method at different time levels.

Fig. 1: *Biomarker Identification Phases.*

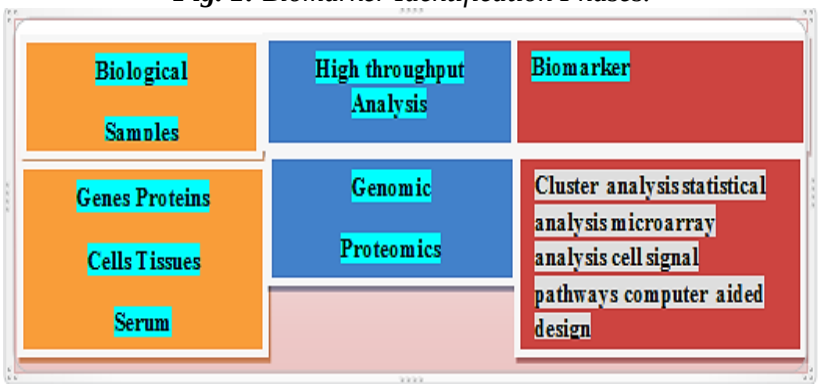


Fig. 2: *For example, given the DNA sequence.*

AAAAACAATTAGCTAACTGGTGATTGTGTGAAGGATGAACTGGATTAGGCCAAGGTGATCAAGAAGAAGA

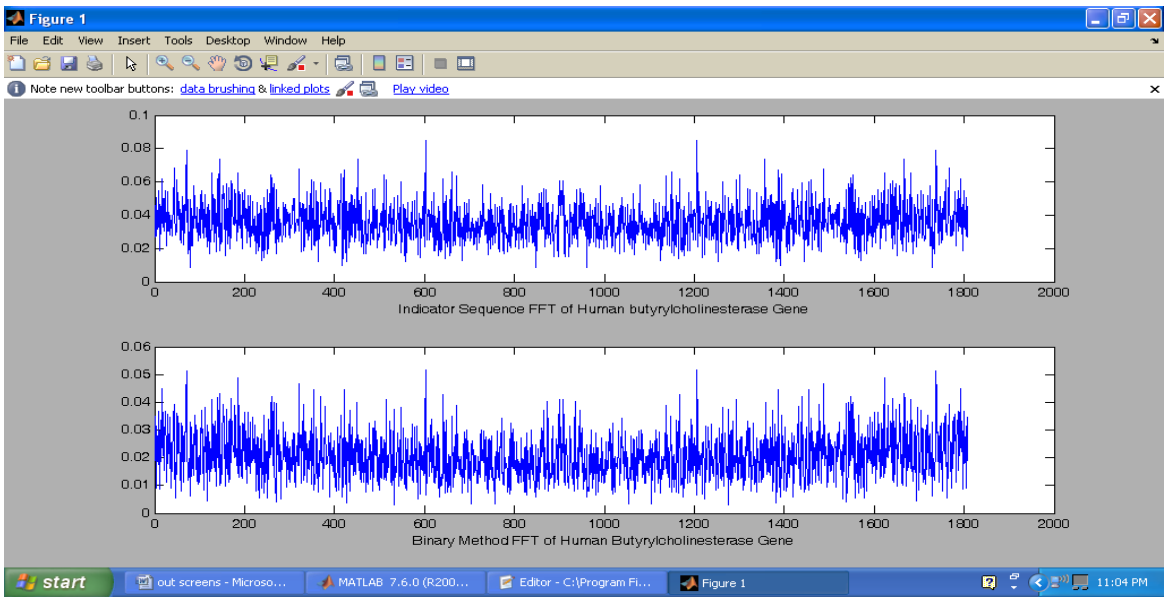
TTGGTAGATTAACGTGGTCAGGAGGTCATGAGAACTTCAAATGAGGCAGTGACCATCAGGAAAAAATTTG

TAAGAAGAATGGTCAGGACCAAATGAGTTTGGTTTGGTCCTGCTGAGTTTGAGGCATATGGTGGAAACTG

CCCAGCTCCCTCCTTCAGAAATGAGACACTCTTTCCCTAGCTGGCCTGGTATAGGCTGTTAATGGCCACC

AGCTGTGTTCCTTTATGGGGCTCGCCCTTGGCTGAAAGGAGCTACAAGGAGTTCATGGGTGACTTTGGCC

AGAGGAGTTGATGAGGAGAGGAAGGTCTGGGG



Method Name	Time Complexity	Length Sequence 384	A residue is converted into a signal Fourier transformation is used
Standard Genomic Signal Method	is More:	0.020868 Seconds	4 x 384 (length of the sequence)
Binary Genomic Signal Method	is reduced	0.013031 Seconds	2 x 384 (length of the sequence

CONCLUSIONS

In this paper the major goal of translational genomics is to discover families of genes whose products (messenger RNA and protein) can be used to classify disease, thereby leading to molecular-based diagnosis and prognosis. We also studied low-level analysis methods for mass spectral data pre-processing. A computational method that combines particle swarm optimization with support vector machines is applied for biomarker selection. The m/z windows selected by the PSO algorithm consist of clearly detectable peaks, which are more likely to represent identifiable proteins, protein fragments or peptides. This is important for our ultimate goal of identifying proteins / peptides that distinguish diabetic patients from healthy individuals. According to the results as shown above we came to a conclusion that PSO is also one of the best tools to solve the optimization problems. PSO parameters are playing a main role in the convergence of all particles. Here we studied the experiment with only Gbest global best) PSO. But there another model called Blest (local best) with two architectures, ring architecture and von-Neumann architecture etc. In initial proteomic discovery studies, three protein species (4.8-, 6.7-, and 13.4-kDa) that were significantly lower in concentration in the CSF from patients with ALS (n = 36) than in normal controls (n = 21) were identified. This work can be further extended to implement the Blest model of PSO that takes the DNA sequence and converted into a genomic signals and identifying the biomarkers and also to establish a medical application DiaBe-Tector that gives a accuracy values for identifying the Hotspots for the drug design.

REFERENCES

1. Dorigo et al. Two Experiments in Embodied Swarm Intelligence. *Web Intelligence and Intelligent Agent Technologies, WI-IAT '09. IEEE/WIC /ACM International Joint Conferences*.2009; 1: 2–3p.
2. Bensalah Karim, Montorsi Francesco, Shariat Shahrokh F. Challenges of Cancer Biomarker Profiling. *European Urology*.2007; 52(6): 1601–1609p.
3. Aziz, K.J. Design controls for development of biomarker-pharmaceuticals. *Journal of Clinical Ligand Assay*. 2008; 31(1-4): 29–48p.
4. Satapathy Suresh Chandra, Murthy JVR. Constrained and Unconstrained function optimization using Swarm Intelligence DW van der Merwe. *AP Engelbrecht, Data Clustering Using Particle Swarm Optimization*. 2003.
5. Kennedy, J., Eberhart, R.C. Particle Swarm Optimization. *In proceedings of the IEEE International Conference on Neural Networks*. 1995; 4: 1942–1948p.
6. F van den Bergh. An Analysis of particle Swarm Optimizers. *PhD Thesis, Department of Computer Science, University of Pretoria, Pretoria, South Africa.* 2002.
7. Satapathy Suresh C et al. Polynomial Neural Swarm Classifier. *in Multimedia University International Symposium on Information and Communication Technology Malaysia.* 24th –25th NOV 2005.
8. Braaten O. et al. The genetic algorithm applied to haplotype data at the LDL receptor locus. *Computer Methods and Programs in Biomedicine*.2000; 61 (1): 1–9p.
9. Dobson CM. The nature and significance of protein folding. In Mechanisms of Protein Folding 2nd ed. Ed. RH Pain. *Frontiers in Molecular Biology series. OxfordUniversity Press: New York, NY.* 2000.