Computational Methods to Locate and Reconstruct Genes for Complexity Reduction in Comparative Genomics

Vidya A.[1], Usha D.[1], Rashma B.M.[1], Deepa Shenoy P.[1], Raja K.B.[1], Venugopal K.R.[1], Iyengar S.S.[2], and Patnaik L.M.[3]

[1] Department of Computer Science and Engineering
University Visvesvaraya College of Engineering
Bangalore University, Bangalore, India
[2] Head of Wireless Sensor Networks and Robotics Research Laboratory
Department of Computer Science
Louisiana State University, Baton Rouge, USA
[3] Vice Chancellor, Defence Institute of Advanced Technology, Pune, India
vidyaananth16@gmail.com

Abstract. Discovering the functions of proteins in living organisms is an important tool for understanding cellular processes. The source data for such analysis are commonly the peptide sequences. Most common algorithms used to compare a pair of nucleotide sequence are Global alignment algorithm (Needleman-Wunch algorithm) or local alignment algorithm (Smith-Waterman algorithm). Analysis of these algorithms show that time complexity required to the above mentioned algorithms is O(mn) and space complexity required is O(mn), where m is size of one sequence and n is size of the other sequence. This is one of the major bottlenecks as most of the sequences are very large. The proposed Coding Region Sequence Analysis(CRSA) algorithm presents a method to reduce both time and space complexity by meaningfully reducing the size of sequences by removing not so significant *e*xons using wavelet transforms. DSP techniques supply a strong basis for regions identification with three-base periodicity.

Keywords: BLAST, Coding Regions, HUMCS3, MGWT, Similarity Search.

1 Introduction

Bioinformatics has been gaining importance for their discoveries in the search for greater understanding of the organisms. With the completion of a host of genome sequencing projects, the amount of available genome data is increasing exponentially [1]. There is a greater need for development of novel methods and techniques of automatic sequencing of large volumes of DNA fragments, prediction of RNA secondary structure and construction of phylogenetic trees. A key

K.R. Venugopal and L.M. Patnaik (Eds.): ICIP 2011, CCIS 157, pp. 252–257, 2011.

step in this process is identification of all genes present in the DNA sequence. Gene Identification is to achieve the annotation in genomics and to look similar to those identified sequences[2], [3]. Different aspects of a sequence search includes search by content, search by signal (also referred as search by site), and search by similarity[4]. Computational methods are necessary to identify genes on the sequenced DNA and knowing the efficiency and reliability, the structure of genes and how they are expressed.

The task involved in the process of identification is the genomic signal processing where in the mapping of the chemical bases of DNA to a number set is achieved. A number of methods have been proposed for gene detection, based on distinctive features of protein-coding sequences. This effective DNA signal is analyzed by the Digital Signal Processing (DSP) concepts and techniques. A properly defined Fourier transform is a powerful predictor of both the existence and the reading frame of protein coding regions in DNA sequences [5].

DNA microarray technology is considered as one of the major step in genomic research due to its parallel processing feature[6]. Wavelet transforms are used to analyse DNA sequences is by taking advantage of multiscale approaches that consists of using small scales to analyze small protein coding regions, and using large scales for larger regions. Wavelet analysis is not suitable in this case since the frequency of the analyzing function varies with the scale parameter, while coding regions with Three Base-pair Periodicity (TBP) presents the same frequency content at different scales[7].

Motivation: Jesus *et* al., [7] presented Modified Gabor-Wavelet Transform (MGWT) for the identification of protein coding regions and it was tuned to analyze periodic signal components and presents the advantage of being independent of the window length. They compared the performance of the MGWT with other methods by using eukaryote data sets. The performance of MGWT is better when compared with all assessed model-independent methods in terms of identification accuracy. It is observed that combined approach of Basic Local Alignment Search Tool(BLAST) and MGWT can reduce time and space complexity by reducing the size of DNA sequence.

Contribution: CRSA algorithm proposed in this paper, reduces the time and space complexity of sequence alignemnt. The intermediate result of the algorithm is a number of homologous sequences. Time and space complexity is reduced by removing non-coding regions of both query and target sequences. The genes were reconstructed and sequence alignment is done for real time analysis.

Organization: The remainder of this paper is organized as follows: Section 2 presents the overview of Literature Survey. Section 3 describes the background. Section 4 describes the proposed method and implementation of reducing the time and space complexity of sequence similarity search. Section 5 presents the results and a comparative assessment of the proposed method with algorithms described in the literature. Concluding remarks are presented in section 6.

2 Related Work

Hwan and Choi [1] proposed the problems associated with gene identification and the prediction of gene structure in DNA sequences. A number of Predictive Methods have been developed to address these problems. Content-based methods rely on the overall, bulk properties of a sequence in making a determination [8]. Characteristics considered here include how often particular codons are used, the periodicity of repeats, and the compositional complexity of the sequence. Because different organisms use synonymous codons with different frequency, such clues can provide insight into determining regions that are more likely to be exons. Lopez-Villasenor et al., [9] proposed the studies consisting of statistical analysis that has the capacity to detect sequence periodicities in DNA.

Tiwari, et al., [5] investigated a Fourier technique based on a distinctive feature of protein-coding regions of DNA sequences, like the existence of short-range correlations in the nucleotide arrangement. The most promising of these is a 1/3 periodicity, which is present in coding sequences. The presence of this periodicity can be seen most directly through the Fourier analysis. Authors focuses on the relative strength of this periodicity, which has been used later in order to form a simple technique to predict genes (with and without introns) in unknown genomic sequences of any organism.

Zhang et al., [6] presented an overview of applications of signal processing techniques for DNA detection, structure prediction, feature extraction and classification of differentially expressed genes.

3 Background

Identification of protein coding regions is an important topic in genomic sequence analysis. Model-Independent methods are not suitable due to their dependence on predefined window length required for a local analysis of a DNA region. Modified Gabor-wavelet Transform (MGWT) proposed by Jesus P et al., [7] for the identification of protein coding regions outperforms all assessed model-independent methods with respect to identification accuracy. This method avoids identification errors but also makes a tool available for detailed exploration of the nucleotide occurrence. A more precise identification of short coding regions is allowed in MGWT. The major problem in large scale data analysis is time and space complexity. With the application of the MGWT sequence length can by reduced by identification of protein coding regions and reconstruction of genes, which greatly enhances the reduction of time and space complexity in genomic data analysis.

4 The Proposed Method

4.1 Problem Statement

Given a raw DNA sequence, the objectives are:
(i) To find a homologous sequence of biological importance.

(ii) To identify a protein coding region.
(iii) To reduce time and space complexity of sequence alignment.

4.2 Algorithm

The raw DNA sequence identified cannot be directly subjected under BLAST. The FASTA format of the DNA sequence has to be fed to BLAST[10].The algorithm proposed Coding Region Sequence Analysis(CRSA) first analyses the sequences withot removing the coding regions. For the second time the algorithm analyses the sequences without non-coding regions.

Table 1. Algorithm: Coding Region Sequence Analysis(CRSA)

Step1: Obtain the FASTA format of the query sequence.
Step2: Find homologous gene for the same by blasting the query sequence.
Step3: Compare query sequence with homologous sequence using Needleman Wunch and Smith-Waterman algorithms using dynamic programming.
Step4: Record the result of comparison of two sequences.
Step5: Remove non-coding region by reconstructing Gene sequence using wavelet transform.
Step6: After removing non-coding regions from both query and target sequence, compare query and target sequences using Needleman Wunch and Smith-Waterman algorithms.
Step7: Record the result of comparison of two sequences after removing non-coding regions.

4.3 Implementation

We focused our study on the analysis of DNA sequences of Eukaryotics. Eukaryotics has been particularily considered for study in the context of coding region identification. For a detailed analysis, we used a sequence Human Chorionic Somatomammotropin hCS-3 gene(HUMCS3)[10]. HUMCS3 is a polypeptide placental hormone. Its function and structure is similar to that of human growth hormone. The target sequence for the further alignment function was identified. Comparison of query sequence with target sequence using Needleman Wunch and Smith-Waterman algorithms using dynamic programming was carried out[11].

For the identification of coding regions on a given DNA sequence the following method is used [7]: (i) Numerically map the DNA sequence to four binary sequences. (ii) For each binary sequence apply the MGWT. (iii)Sequences spectra has to be projected onto the position axis and (iv) Thresholding the projection coefficients for location of the edges among coding regions. After removing non-coding regions from both query and target sequence, comparison of query and target sequences was done using Needleman Wunch and Smith-Waterman algorithms[10].

5 Results and Discussions

The FASTA format of the raw DNA sequence HUMCS3, has been subjected under BLAST[10] for the identification of the homologus sequence, which has resulted in a number of sequences. The homologous sequence obtained from BLAST is in good agreement with Homo sapiens placental lactogen hormone precursor (CSH1) gene, complete cds(HUMPLA). It modifies the metabolic state of the mother during pregnancy to supply the energy to the fetus. HUMPLA is considered as the target sequence with a 97% similarity against the query sequence. The two sequences have been aligned for similarity search[11]. The MGWT method extracts all *e*xons and *i*ntrons. The non-coding regions were removed using MGWT[7]. The protein coding regions of both query and target sequences are as shown in the Table 2.

The identity and similarity of query sequence and the target sequence was 90.6% before removal of the non-coding regions. The score of similarity search is found to be 13213.5. The Smith Waterman approach has minimum number of gaps when compared to Needleman Wunch approach. After removing the non-coding regions, the reconstructed genes are again aligned for similarity search[11]. The result obtained is depicted in Table 3.

The percentage of identity is found to be 90.0% in Needleman and Wunch algorithm, whereas it is 98.0% in Smith-Waterman approach. The percentage of similarity is 90.3% in Needleman and Wunch algorithm and it is found to be 98.0% in Smith-Waterman approach. The number of gaps were reduced to 3 in case of Smith waterman approach. The score after removing the non-coding regions is found to be 965. From the results obtained we observe that the similairty of the two sequences is almost equal in both the approaches. The time and space complexity of the sequence alignment is reduced in this approach.

Table 2. Protein Coding Region of HUMCS3 and HUMPLA

HUMCS3		HUMPLA	
Base Pair Count	No.of Base Pairs	Base Pair Count	No.of Base Pairs
1:1240-1261	22bp	1:1224-1273	50bp
2:1380-1582	203bp	2:1385-1595	211bp
3:1838-2029	192bp	3:1830-2019	190bp

Table 3. Results Obtained Before and After Removal of Non-Coding Regions

Factors	Before Removal		After Removal	
	HUMCS3	HUMPLA	HUMCS3	HUMPLA
	Needle	Water	Needle	Water
Length	2979	2754	217	200
Identity	2700/2979 (90.6%)	2700/2754 (98.0%)	196/217 (90.0%)	196/200 (98.0%)
Similarity	2700/2979 (90.6%)	2700/2754 (98.0%)	196/217 (90.3%)	196/200 (98.0%)
Gaps	251/2979 (8.4%)	26/2754 (0.9%)	20/217 (9.2%)	3/200 (1.5%)
Score	13213.5	3213.5	965.0	965.0

6 Conclusions

The role of signal processing in genomics is quite important. In the present work, a wavelet transformed method is adopted for finding coding regions in a query sequence (HUMCS3). Homologous sequence is found using BLAST. Homologous sequence HUMPLA is in good agreement with HUMCS3. A modified Gabor Wavelet Transform is adopted to remove the non-coding regions. The time and space complexity of sequence alignment is greatly reduced after removing non-coding regions.

References

1. Do, J.H., Choi, D.K.: Computational Approaches to Gene Prediction. The J. Microbiology 44(2), 137–144 (2006)
2. Zhang, M.Q.: Computational Prediction of Eukaryotic Protein Coding Genes. Nature Rev. Genetics 3(9), 698–709 (2002)
3. Dougherty, E.R., Shmulevich, I., Chen, J., Jane Wang, Z.: Genomic Signal Processing and Statistics, vol. 2. Hindawi Publishing Corp., (2005)
4. Blanco, E., Guigo, R.: Predictive Methods using DNA Sequences, Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd edn. John Wiley and Sons, Inc., Chichester (2004)
5. Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., Ramaswamy, R.: Prediction of Probable Genes by Fourier Analysis of Genomic Sequences. Bioinformatics 13(3), 263–270 (1997)
6. Zhang, X., Chen, F., Zhang, Y., Agner, S.C., Akay, M., Lu, Z., Waye, M. M. Y., Tsui, S. K.: Signal Processing Techniques in Genomic Engineering. In: IEEE, pp. 1822–1833 (2002)
7. Mena-Chalco, J.P., Carrer, H., Zena,Y., Cesar Jr., R.M.: Identification of Protein Coding Regions using the Modified Gabor-Wavelet Transform. IEEE/AMC Transactions on Computational Biology and Bioinformatics 5(2) (2008)
8. Baxevanis, A.D., Francis Ouellette, F.B.: Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 2nd edn. A John Wiley and Sons, Inc., Chichester (2001)
9. Lopez-Villasenor, I., Jose, M.V., Sanchez, J.: Three-Base Periodicity Patterns and Self-Similarity in Whole Bacterial Chromosomes. Biochemical and Biophysical Research Comm. 325(2), 467–478 (2004)
10. National Center for Biotechnology Information, `http://ncbi.nlm.nih.gov`
11. `http://www.ebi.ac.uk/Tools/Psa/`