Computational Methods in Biological Sequence Analysis for Complexity Reduction

Vidya A\textsuperscript{a}, Usha D\textsuperscript{a}, Rashma B M\textsuperscript{a}, K R Venugopal\textsuperscript{a} and L M Patnaik\textsuperscript{b}

\textsuperscript{a}Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore-560 001, India, Contact: vidyaananth16@gmail.com

\textsuperscript{b}Vice Chancellor, Defence Institute of Advanced Technology (Deemed University), Pune, India.

Exploring the functions of proteins is an important tool for understanding cellular processes in living organism. The general source data for analysis are the peptide sequences. The most common algorithms used to compare a pair of nucleotide sequence are Global alignment algorithm or local alignment 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 exons using wavelet transforms.

DSP techniques supply a strong basis for regions identification with three-base periodicity.

\textbf{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 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, and search by similarity [4]. Search by content searches for DNA segments with specific properties like the composition of the nucleotides, nucleotide frequency, composition of codons. The search by signal and search by similarity methods depends on the known database which is used to train a classifier [5]. In a different classifcation of coding region identification, the methods have been divided into model dependent and model independent methods. Model dependent methods are based on a prior information where as model independent methods do not depend on prior information, instead it depends on the detection of the periodic regions [6].

Computational methods are necessary to identify genes on the sequenced DNA and knowing the efficiency and reliability, the structure of genes and when and how they are expressed. 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 detec-
Table 6
Results Obtained After Removal of Non-Coding Regions

<table>
<thead>
<tr>
<th>Factors</th>
<th>HUMCS3 V/S HUMPLA After Removal</th>
<th>HUMCS3 V/S HUMPLA After Removal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Needle</td>
<td>Water</td>
</tr>
<tr>
<td>Length</td>
<td>217</td>
<td>200</td>
</tr>
<tr>
<td>Identity</td>
<td>196/217 (90.0%)</td>
<td>196/200 (98.0%)</td>
</tr>
<tr>
<td>Similarity</td>
<td>196/217 (90.3%)</td>
<td>196/200 (98.0%)</td>
</tr>
<tr>
<td>Gaps</td>
<td>20/217 (9.2%)</td>
<td>3/200 (1.5%)</td>
</tr>
<tr>
<td>Score</td>
<td>965.0</td>
<td>965.0</td>
</tr>
</tbody>
</table>

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 similarity of the two sequences is almost equal in both the approaches. The time and space complexity of the sequence alignment is greatly reduced by considering only the coding regions.

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

12. Andreas D Baxevanis, Francis Ouellette, B F.


16. [http://www.ebi.ac.uk/Tools/Psa/](http://www.ebi.ac.uk/Tools/Psa/).

**Vidya A** is an Assistant Professor in the Department of Information Science and Engineering at Vivekananda Institute of Technology Bangalore, India. She received her Bachelor of Engineering and Masters of Engineering degrees in Computer Science and Engineering from Bangalore University, Bangalore. She is presently pursuing her Ph.D programme in the area of Data mining in Jawaharlal Nehru Technological University, Hyderabad. Her research interests are in the area of bioinformatics, data mining and Image processing.

**Usha D** received her BE degree from Visvesvaraya Technological University, Belgaum and ME degree from Bangalore University. Presently working as Lecturer in PESIT, Bangalore. Her research interests are in the area of Computational Biology, Data Mining and Embedded systems and design.

**Rashma B M** received her BE degree from Visvesvaraya Technological University, Belgaum and ME degree from Bangalore University. Currently working as a Lecturer in PESSE, Bangalore. Her research interests include Computational Biology, Pattern Recognition, and Data Mining.

**Venugopal K R** is currently the Principal, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering. He received his Masters degree in Computer Science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D. in Economics from Bangalore University and Ph.D. in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored 31 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ and Digital Circuits and Systems etc.. During his three decades of service at UVCE he has over 250 research papers to his credit. His research interests include Computer Networks, Wireless Sensor Networks, Parallel and Distributed Systems, Digital Signal Processing and Data Mining.

**L M Patnaik** is a Vice Chancellor, Defense Institute of Advanced Technology, Pune, India. He was a Professor since 1986 with the Department of Computer Science and Automation, Indian Institute of Science, Bangalore. During the past 35 years of his service at the Institute he has over 700 research publications in refereed International Journals and refereed International Conference Proceedings. He is a Fellow of all the four leading Science and Engineering Academies in India; Fellow of the IEEE and the Academy of Science for the Developing World. He has received twenty national and international awards; notable among them is the IEEE Technical Achievement Award for his significant contributions to High Performance Computing and Soft Computing. His areas of research interest have been Parallel and Distributed Computing, Mobile Computing, CAD for VLSI circuits, Soft Computing and Computational Neuroscience.