IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol-6, Special Issue-4, May 2018 E-ISSN: 2347-2693

A Survey on Computational Algorithms For Biological Data Analysis

M.Muthu Lakshmi[1], G.Murugeswari[2]

[1,2]Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India

**Corresponding Author: muthulakshmi311292@gmail.com, Mob: 9952627205*

Available online at: www.ijcseonline.org

***Abstract*—**Bioinformatics is an interdisciplinary field that uses the information technology algorithms for biological data analysis. Many tools and techniques have been investigated by the researchers for biological data interpretation, analysis and prediction. In accordance with the latest statistics, biological sequence analysis is one of the emerging areas in the field of Bioinformatics. In this paper, a survey on computational algorithms of Bioinformatics has been made. We analyzed the contributions made by the computer researchers for biological sequence analysis and survey is presented on various categories such as biological sequencing, alignment, compression and encoding, feature extraction, clustering and classification. The objective of the paper is to provide a deep understanding and knowledge regarding the existing computer algorithms used for biological data analysis and to identify the research areas for computer researchers in the field of bioinformatics.

***Keywords*—**Bioinformatics, Biological Sequences, DNA, RNA, Protein

I. Introduction

Bioinformatics is a hybrid field that brings together the knowledge of information science and knowledge of biology. This field is playing an important role in genetic research. A gene or genetic regulatory network is a collection of molecular regulators which interact with each other substances in the cell to govern the gene expression levels. The regulator can be DNA(DeoxyRibo Nucleic Acid), RNA (Ribo Nucleic Acid), protein and complexes of these.

Human body is made up of trillion of cells. According to human genome theory, the number of genes in each cell is approximately 20,000. Gene is the hereditary unit that inherits features from ancestors in a living organism. Every organism has different genes corresponding to different characters. Some are visible characters like height and colour. Some are invisible characters like biological process inside the body. Genes are built from long molecules called DNA which lies in the chromosome. The process of transcription of DNA to RNA and translation of RNA to protein is depicted in Figure 1.

DNA
RNA
Proteins

Figure1. Biological changes in sequences

- DNA carries genetic information in the form of building blocks A, C, G, Twhere A stands forAdenine, C for Cytosine, G for Guanine and T for Thymine.
- RNA is the intermediate stage in the process of genetic information transformation.
- Proteins comprised of amino acids, perform most cell functions.

In order to study the biological processes and to increase the understanding of organisms, researchers have been engaged in analysing and interpreting various types of biological data, one of which is the biological sequences. The term biological sequence refers to DNA sequences, RNA sequences and Protein sequences. A biological sequence is a single, continuous molecule of nucleic acid or protein. Sequence analysis is the process of subjecting a DNA, RNA or peptide sequence to any of the wide range of analytical methods to understand its features, functions, structure or evolution. DNA sequence is used to predict the genetic diseases, age prediction and cancer risk prediction in organisms. DNA finger printing helps to identify the child's parents. Protein sequence helps to understand how the protein functions in the system and also to identify the protein deficiency diseases.

Computers are used for enormous data storage, retrieval and analysis. Researchers proposed few computer techniques for biological sequence encoding, feature extraction, clustering, classification and similarity findings. The length of biological sequence is one of the most challenging issues faced by the researchers. Because of its length, the storage space and the processing timeis also seems complicated.

The objective of this work is to analyse the existing computer techniques which have been used by the researchers for biological sequence analysis and to find the research scope in the field of bioinformatics for computer researchers.

Rest of the paper is organized as follows, Section II contains review of related works, Section III describes the resources which include some tools, databases and websites used for bioinformatics research and Section IV concludes the survey with future scope.

II. LITERATURE REVIEW

We reviewed many computational biology research papers and survey report is presented under the following categories.

A. BIOLOGICAL SEQUENCING

Erik Pettersson et al. [29] analyzed few sequencing technologies in terms of processing speed and cost. The author found that miniaturized and parallelized platforms allow lower sample and template consumption which helps to increase the speed and reduce the computation cost. It is also investigated that massively parallel systems provide significantly improved throughput over Sanger systems (First generation sequencing systems) and single molecular approaches improve the performance.

James M. Heather and Benjamin chain [16] presented an approach which measures the changes moving from sequencing small number of bases to millions of bases. It also describes the challenges faced from detection of coding sequence of a single gene to rapid and widely available whole genome sequencing. It provides a brief history of different generations of sequencing technology.

Michel C. Wendl et al. [40] proposed a software framework “Hands-off: analysis and handling of DNA data”. The authors also discussed a number of critical software algorithms, components and the method of which they have been woven into a framework for hands-off processing.The Drawback of the work is that the proposed work needs human intervention.

Yang Chen and Jinglu Hu [8] proposed a method for DNA sequencing by hybridization technique based on constructive heuristic algorithm. The experiments were carried out using benchmark instance setsand proved that the proposed method achieved higher accuracy for reconstruction of long DNA sequences.

Jacek Blazewicz et al. [4] presented a graph models used for sequencing by hybridization and also discussed their properties and connection between them. The author explained how the graph models evolved to adapt to next generation sequencing.Practical comparisons of various DNA de novo assembly tools are also provided.

B. ALIGNMENT OF BIOLOGICAL SEQUENCES:

Zheng Zhang et al. [43] proposed a greedy algorithm for aligning DNA sequences. The genomic sequence of Mycobacterium Tuberculosis strain H37Rv is taken for the purpose of alignment. Initially, an X-Drop algorithm is formulated which uses dynamic programming approach. After that, a greedy X-Drop algorithm is developed for alignment of sequences. Results of both the algorithm are compared and the greedy algorithm works better than dynamic programming algorithm in terms of time.The author leaves the development of an efficient alignment algorithm for symbol dependent match scores as an open problem.

Jeong-Hyeon Choi et al. [9] proposed a genome alignment algorithm namely GAME (Genome Alignment by Match Extension) based on MEM (Maximal Exact Match) anchors. An experiment is performed using 10 microbial genomes of M.genitalium. Performance of GAME is compared with the existing algorithms such as BLASTZ and MUMmer. As a result, GAME performs better in terms of runtime and it also achieves high quality sequence alignment.

Done Stajanov and Aleksandra Mileva [35] presented a survey on pair wise local and global sequence alignment algorithms together with comparative analysis. The algorithms such as Needleman-Wunsch, Hirschberg, Smith-waterman, Waterman-eggert, Huang-Miller, FASTA, MUMmer, AVID, SPA, VMatch are compared in terms of its Characteristics, Time complexity and Space complexity.Several algorithms which haveboth time and storage complexities are not included.

C. COMPRESSION AND ENCODING SCHEME OF BIOLOGICAL SEQUENCES

Stephane Grumbach and Fariza Tahi [13] proposed a lossless compression algorithm based on regularities such as the presence of palindromes in DNA sequence. It is observed by the authors that the results are not satisfactory, but far beyond the classical algorithms.

Sebastian Wandelt and Ulf Leser [38] proposed a general open source framework namely FRESCO (Framework for Referential Sequence Compression) to compress large amount of biological sequence data. The proposed method FRESCO is compared with the standard existing referential compression algorithms. As a result, FRESCO works much better than the existing algorithms in terms of compression speed. Several techniques have been proposed to increase the compression ratio by selecting a good reference sequence and by rewriting a reference sequence. In addition, a new way of boosting the compression ratio is achieved by applying referential compression to already referentially compressed files. This method is referred as second order compression.

FASTQ format is a text based format to store the biological sequences with their corresponding quality scores. Subrata Saha and Sanguthevar Rajasekaran [33] proposed a novel algorithm namely FQC (FASTQ Compressor) for compressing FASTQ files. In the proposed work, the FASTQ file compression problem is solved by using three different algorithms namely RFRC (Reference Free Reads Compressor), RFQSC (Reference Free Quality Scores Compressor) and MDC (Metadata Compressor). All of these three algorithms are combined into a single algorithm called FQC algorithm. Implementation is done using some real time datasets. Experimental results show that the proposed algorithm is effective and efficient than the existing algorithms in the domain of FASTQ file compression.

Marius Nicolae et al. [27] introduced a lossless non-reference based FASTQ compression algorithm namely LFQC (Lossless FASTQ Compressor). The proposed algorithm is compared with few existing compression algorithms and better compression ratio is achieved for LS454 and SOLiD Datasets. The algorithm is implemented in Ruby programming language, if the implementation is done using C or C++ then the speed of the algorithm may increases.

Xin Chen et al. [7] proposed a lossless compression algorithm namely GenCompress for genetic sequences. It is reported that the approximate repeats are one of the main hidden regularities in DNA sequences. GenCompress is developed based on approximate repeats and tested on standard benchmark DNA sequences. The experimental results of the proposed algorithm are compared with Biocompress-2 and Cfact and better performance is proved.Xin Chen presented another compression algorithm called DNA Compress which improves the running time of all previously proposed DNA compression algorithms.

In the year of 2009, Raffaele Giancarlo et al. [12] presented a review report on key areas of bioinformatics and computational biology where compression techniques have been used. In continuation with the review report, the authors presented a companion review in 2012 to explore the computational methods involved in the use of data compression in Bioinformatics.

Reference based compression method was invented by Markus Hsi-Yang Fritz et al. [11] to compress the DNA sequences for storage. The storage of quality scores and unaligned sequences are adjustable for different experiments to conserve the data information or to minimize the storage cost.The limitation is that the algorithm is implemented using python programming language and this prototype works only with BAM inputs.

Armando J. Pinho et al. [30] demonstrated the ability of finite context models for DNA Sequence compression. The authors also proposed a compression method based on eight finite-context models and tested the method on total of 2,338 bacterial genome sequences. It provides better results than XM-DNA coding algorithm.

An expert model (XM) and an algorithm have been introduced by Minh Duc Cao et al. [5] for biological sequence compression that makes use of statistical properties and repetition within the sequences. In this work, the probability distribution of the next symbol in the sequence is estimated. The symbols are encoded using arithmetic coding.

A compression algorithm using dynamic programming “DNA Pack” was proposed by Behshad Behzadi and Fabrice Le Fessant[2]. The DNA Pack is based on dynamic programming approach and so it provides better compression ratiothan few existing compression algorithms.

Gergely Korodi and Ioan Tabus [22] introduced a compression method by combining encoding using normalized maximum likelihood model and encoding by first order context coding.The proposed work has been tested with recently published human genome data.

Waibhav Tembe et al. [37] proposed Genomic Squeez (G-SQZ), a Huffman coding based sequencing scheme that compresses data without altering the relative order. The proposed method improves the compression ratio on benchmark datasets and the results are compared with standard compression tools namely gzip and bzip2.

Faraz Hach et al. [15] proposed SCALCE (Sequence Compression Algorithms using Locally Consistent Encoding), a boosting scheme for the compression of biological sequences. It isdeveloped based on Locally Consistent Parsing technique, which reorganizes the reads so that highercompression speed andcompression rateare obtained. Experimental results indicate that SCALCEcan provide upto 3.34% improvement in compression rate and 1.26%improvement in running time.

Carl Kingsford and Rob Patro [21] presented a Path encoding approach for compression to reduce the difficulty of managing large scale sequencing data. The proposed method draws a connection between storing paths in de Bruijn graphs and context-dependent arithmetic coding. This method is able to encode RNA sequence reads using 3-11% of the space of the sequence in raw FASTA files. The experimental results are compared with various compression methods such as SCLACE, fastqz and CRAM and the ability of the method is proved in term of compression size.Path encoding does not allow random access to records in compressed file.

D. FEARTURE EXTRACTION OF BIOLOGICAL SEQUENCES

QingZhou and Jun S. Liu [45] presented a systematic study of predictive modeling approaches to the TF-DNA binding problem. In these approaches, a statistical relationship between genomic sequences and gene expression is inferred through a regression framework and influential sequence features are identified by variable selection. Features such as GC content, average CVs (Conservation Score), sequence length, counts of k-mers and motif features are extracted. The methods such as Stepwise linear regression, Multivariate adaptive regression splines, Neuralnetworks, SVM, Boosting and Bayesian Additive Regression Trees(BART) are applied to both simulated datasets and two whole genome CHIP-chip datasets. As a result, BART and Boosting proved the best and the most robust performance among all other methods.

Online hierarchical feature extraction algorithm was introduced by Mehdi Kchouk and Faouzi Mhamdi [19] for classification of protein sequences. N-grams method has been adapted to extract the feature of variable sizes.The use of N-grams allows extracting good features. The features extraction method is tested with supervised learning algorithm for classification of real protein banks data.

Xianwen Ren et al. [32] proposed a feature extraction method called ipcc(iterative pearson correlation coefficients).ipcc is also used for disease class discovery and prediction based on high throughput gene expression profiles such as RNA Sequence.This proposed feature extraction method is expected to be a useful tool for the development of clinical diagnosis.

Jason T.L.Wang [39] proposed a technique to extract features from protein sequences. The author initially encodes the protein sequences and then extracts features such as global similarity and local similarity from the sequences. The extracted features are given as input to the BNN (Bayesian Neural Network) classifier for classification of protein sequences.The proposed BNN classifier is compared with other three classifiers namely BLAST,SAM and SAM-T99. The result shows that BNN classifier takes more time in training phase but once it is trained, it provides results faster than other classifiers.

Zena M. Hira and Duncan F. Gillies [17] performed a review of feature selection and feature extraction methods used on microarray cancer data. Different feature selection and feature extraction methods were described and compared. The author also discussed the advantages and disadvantages of each method.

RabieSaidi et al. [34] proposed an encoding method that uses amino-acid substitution matrices to define similarity between motifs. SVM classifier is used for protein sequence classification. In order to prove the efficiency, the proposed approach is compared with several encoding methods which use some machine learning classifiers.

E. CLUSTERING OF BIOLOGICAL SEQUENCES

Abdellali Kelil et al. [20] proposed a similarity measure based on matching amino acid subsequences. A measure named SMS (Substitution Matching Similarity) is especially designed for application to non-aligned protein sequences. It leads to the development of new alignment free algorithm named CLUSS, for clustering protein families. To show the effectiveness of the proposed method, an extensive clustering on COG database is performed. The experimental results of the proposed method CLUSS is compared with the existing methods such as BLAST, TRIBE-MCL and gSpc. As a result, CLUSS yields a Q-measure value of 87.09%. Also, CLUSS works well on both aligned and non-aligned protein sequences.

A graph based clustering method was proposed by Hideya Kawaji et al.[18] to cluster protein sequences into families. The proposed approach formulates sequence clustering problem as a kind of graph partitioning problem in a weighted linkage graph, which vertices correspond to sequences and edges correspond to higher similarities than given threshold and are weighted by their similarities. The effectiveness of the proposed method is proven by comparing it with Interpro families in all mouse proteins in SWISS-PROT. As a result, by using the proposed method 77% of proteins in Interpro families are classified into appropriate clusters.

Valerie Guralnik and George Karypis [14] presented an approach for sequence clustering which uses k-means based clustering algorithm. Initially the similarity between the sequences is measured and then clustering is done based on the features and similarity of the sequences. Comparison of feature based clustering and similarity based clustering is also made. The feature based approach acquires better accuracy than similarity based approach.

J.D.Parsons et al. [28] proposed an algorithm namely ICAtools for clustering of cDNA sequences. The proposed algorithm is comprised of three separate programs namely ICAtool, ICAprint and ICAstats. ICAtools are used to cluster similar sequences together. The limitation is that the current version of ICA tool does not support repeat clustering of large datasets.

A clustering of DNA sequences by feature vectors was proposed by Libin Liu et al. [24]. In this work, a DNA sequenceis represented as a point in twelve dimensional spacesand a twelve-dimensional vector is derived. Experiments were carried out on myoglobin, b-globin, histone-4, lysozyme and rhodopsin families.

Weizhong Li and Adam Godzik [23] derived few programs based on Cdhit algorithm for clustering of large sets of protein or nucleotide sequences. The proposed programs include cd-hit-2d, cd-hit-est and cd-hit-est-2d.Cd-hit-2d compares two protein datasets and reports similar matches between them, cdhit-est clusters the DNA/RNA database and cd-hit-est-2d compares two nucleotide datasets.All these proposed programs can handle huge number of sequences and can process hundreds of times faster than BLAST, a popular sequence comparison tool.

F. CLASSIFICATION OF BIOLOGICAL SEQUENCES

Sanghamitra Bandyopadhyay [1] proposed a hybrid technique for classifying amino acid sequences into different super families. This method combines the techniques of feature extraction, fuzzy clustering and NN classification. The results are reported on three superfamily classes namely globin, trypsin and ras. The experimental results are also compared with that of BLAST, NN and FCM methods. It is found that the time requirement for classification using the proposed method and FCM is significantly lower as compared to the other methods.

A web based system CLASSSEQ was developed by Kwangmin Choi et al. [10] for analysis and comparison of uncharacterized protein sequences against multiple genomes. The user sequences are combined with protein sequences from theuser specified genomes and then clustered using BAG clustering algorithm. The pre-computed genome to genome Pair wise Comparison DataBase (PCDB) makes the service faster. CLASSSEQ allows the users to select any combination of genomes and then perform clustering analysis by using BAG. CLASSSEQ uses several in-house tools namely BAG, MCGS and operon viz.

Golan Yona et al. [41] discussed about the ProtoMap site which is used for automatic classification of protein sequences in SWISS-PROT database. The classification is based on pairwise similarities among the protein sequences.

Henrik Stranneheim et al. [36] introduced a FACs (Fast and Accurate Classification of sequences) that can accurately and rapidly classify the sequences. Experiment was done on metagenome dataset and the results show that the proposed method achieved more accuracy and performs faster when compared with tools such as BLAT and SSAHA2.

G. SOME OTHER RESEARCHES IN BIOLOGICAL SEQUENCES

Motif discovery

Definition of Motif

Sequence motif is nucleotide or amino-acid sequence pattern that is widespread and has biological significance. Qiang Yu et al. [42] proposed a motif discovery algorithm named MCES which identifies motifs by mining and combining emerging substrings. To handle larger data sets, a Map reduction based strategy is designed to mine emerging substrings distributed. Experimental results shows that MCES is able to identify motifs efficiently and effectively in thousands to millions of input sequences and it provides better identification accuracy than the competing algorithms namely CisFinder.

Ramanujam and Padmavathi [31] proposed an algorithm namely CFMD (Constraint Frequent Motif Detection) for extracting both contiguous, non-contiguous patterns of short or long sequences of any length in biological database. CFMD combines data mining techniques such as TRIE like Frequent Pattern (FP-TREE) in constructing the patterns. The performance of the proposed approach is proved using both real and synthetic datasets and the results show that CFMD is fast and scalable to extract patterns.

Repeat Finding

Gary Benson [3] presented an algorithm namely TRF (Tandem Repeat Finder) for finding repeats which works without the need to specify either the pattern or pattern size. The algorithm uses probabilistic models for finding repeats. The algorithm is analyzed and implemented upon four sequences such as human frataxin gene sequences, human βT cell receptor locus sequences and two yeast chromosomes.The limitation is that some repeats which undergo several mutations are missed by TRF.

Hongxia Zhou[44] presented a spectral method based on the AR (Auto Regressive) model to detect tandem repeats in DNA sequences. In this method, the spectrogram of a DNA sequence is analyzed based on the autoregressive model. Then, significant peaks in the spectrogram are selected and the corresponding regions in the DNA sequence are analyzed to find the tandem repeats. Experiment results show that the proposed method has a superior performance when compared with TRF.But the drawback is that sometimes, the method is too weak to detect the spectrum magnitude for smallest frequency.

Pattern Similarity recognition

Piero Montanari et al. [26] discovered an optimized pattern search algorithm to find pattern within a large set of genomic sequences, which are similar to the given pattern. The pattern similarity is implemented using dynamic programming and is enhanced with efficient window-based-technique.

Lei Chen et al. [6] proposed derivative Boyer-Moore (d-BM), a pattern matching algorithm in DNA sequences. This algorithm has been developed based on the Boyer-Moore method.Compression of DNA sequences and patterns is carried out by using 2 bits to represent each of A,T,C,G characters.

Mayilvaganam and Rajamani [25] implemented a system using Hidden Markov Model (HMM) which is mostly used in pattern recognition domain. The work focused and analyzed about performance of DNA gene liver cancer database and normal liver cell data set from NCBI database. The proposed HMM System is validated with two different nucleotide values to analyze the performance.

III. RESOURCES

Some of the useful tools, software, databases and websites used in the analysis of biological sequences are presented in this section.

A. Tools and software

Few tools and software that are used for biological sequence analysis such as alignment, compression, comparison, clustering and classification are provided inTable 1.

Table 1. Computational Tools and Softwarefor sequence analysis

Process	Tools and Software
Sequence Alignment	Clustal W2, DIALIGN, LALIGN, MAFFT, MUSCLE, TCoffee, BLASTN, BLASTX, TBLASTX, DotLet, BALSA etc.
Sequence Comparison	BLAST,BLASTP,FASTA,FASTP,TFASTA,LFASTA,RDF2, QUAST etc.
Sequence Compression	gzip and bzip, fqzcomp, DSRC2,BEETL etc.
Sequence Clustering	DASP3, CLAP, Motifcluster, MeshClust, DNACLUST and fMLC etc.
Sequence Classification	ProtoMap, CASTOR, SDT, SPiCE, UPROC etc

B. Databases

Table 2 contains the details of Databases for Biological sequence data.

Table 2. Databases for Biological sequences

Database	Sequence Type	Description
NCBI	Nucleotide sequences and Protein sequences	National Center for Biotechnology Information is the most popular database which provides access to biomedical and genomic information.
GenBank	Nucleotide sequence	One of the largest public sequence databases
DDBJ	Nucleotide sequence	DNA Databank of Japan
EMBL	Nucleotide sequence	European molecular Biology Laboratory
MGBD	Nucleotide sequence	Mouse Genome Database
GSX	Nucleotide sequence	Mouse Gene Expression Database
NDB	Nucleotide sequence	Nucleic Acid Database
Rfam	Nucleotide sequence	RNA family Database
SWISS-PROT	Protein sequence	Swiss Institute for Bioinformatics and European Bioinformatics Institute
TrEMBL	Protein sequence	Annotated supplement to SWISS-PROT
PIR	Protein sequence	Protein Information Resource
Pfam	Protein sequence	Database consists of large collection of protein families

C. URLs

Some of the useful websites for Biological tools and databases are as follows:

https://www.expasy.org/ - Bioinformatics resource portal which provides access to scientific databases and software tools.

http://www.cellbiol.com/sequence_tools.php - Sequence analysis tools and databases for molecular biology and bioinformatics.

http://www.internationalgenome.org/ - Provides access to software tools and also allows download of the genome sequences.

https://molbiol-tools.ca/Genomics.htm - Genomics portal which provides a large number of tools for genomic analysis.

https://omictools.com/genomics2-category - Provides a complete set of software tools and databases used in various stages of genomic processing.

IV. CONCLUSION

This paper presents a survey on computational algorithms applied on biological data. Various algorithms, techniques, tools, software, bioinformatics databases and related websites are presented in this survey.It is found that the following research areas are found for the computer researchers on biological sequences in the field of Bioinformatics.

1. Compression and encoding
2. Sequence Alignment
3. Pattern Matching
4. Gene Recognition/ Promoter Recognition
5. Similarity measure
6. Clustering
7. Classification
8. Feature Extraction /Selection
9. Repeat finding/ Motif Discovery

Since we focus on feature selection and encoding techniques, the issues and challenges are being analyzed in these domains. It is planned to propose a novel technique for feature selection and encoding schemes for biological data sequences in future.

V. REFERENCES

[1] Bandyopadhyay, Sanghamitra. "An efficient technique for superfamily classification of amino acid sequences: feature extraction, fuzzy clustering and prototype selection." Fuzzy Sets and Systems 152.1 (2005): 5-16.

[2] Behzadi, Behshad, and Fabrice Le Fessant. "DNA compression challenge revisited: a dynamic programming approach." Annual Symposium on Combinatorial Pattern Matching. Springer, Berlin, Heidelberg, 2005.

[3] Benson, Gary. "Tandem repeats finder: a program to analyze DNA sequences." Nucleic acids research 27.2 (1999): 573.

[4] Blazewicz, Jacek, Marta Kasprzak, Michal Kierzynka, Wojciech Frohmberg, Aleksandra Swiercz, Pawel Wojciechowski, and PiotrZurkowski. "Graph algorithms for DNA sequencing–origins, current models and the future." European Journal of Operational Research 264, no. 3 (2018): 799-812.

[5] Cao, Minh Duc, Trevor I. Dix, Lloyd Allison, and Chris Mears. "A simple statistical algorithm for biological sequence compression."

In Data Compression Conference, 2007. DCC'07, pp. 43-52. IEEE, 2007.

[6] Chen, Lei, Shiyong Lu, and Jeffrey Ram. "Compressed pattern matching in DNA sequences." In Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE, pp. 62-68. IEEE, 2004.

[7] Chen, Xin, Sam Kwong, and Ming Li. "A compression algorithm for DNA sequences." IEEE Engineering in Medicine and biology Magazine 20.4 (2001): 61-66.

[8] Chen, Yang, and Jinglu Hu. "Accurate reconstruction for DNA sequencing by hybridization based on a constructive heuristic." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 8.4 (2011): 1134-1140.

[9] Choi, Jeong-Hyeon, Hwan-Gue Cho, and Sun Kim. "GAME: a simple and efficient whole genome alignment method using maximal exact match filtering." Computational Biology and Chemistry 29, no. 3 (2005): 244-253.

[10] Choi, Kwangmin, Youngik Yang, and Sun Kim. "CLASSEQ: Classification of Sequences via Comparative Analysis of Multiple Genomes." Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on. IEEE, 2007.

[11] Fritz, Markus Hsi-Yang, Rasko Leinonen, Guy Cochrane, and Ewan Birney. "Efficient storage of high throughput DNA sequencing data using reference-based compression." Genome research 21, no. 5 (2011): 734-740.

[12] Giancarlo, Raffaele, Davide Scaturro, and Filippo Utro. "Textual data compression in computational biology: a synopsis." Bioinformatics 25.13 (2009): 1575-1586.

[13] Grumbach, Stéphane, and FarizaTahi. "Compression of DNA sequences." In Data Compression Conference, 1993. DCC'93., pp. 340-350. IEEE, 1993.

[14] Guralnik, Valerie, and George Karypis. "A scalable algorithm for clustering sequential data." Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE, 2001.

[15] Hach, Faraz, Ibrahim Numanagić, Can Alkan, and S. CenkSahinalp. "SCALCE: boosting sequence compression algorithms using locally consistent encoding." Bioinformatics28, no. 23 (2012): 3051-3057.

[16] Heather, James M., and Benjamin Chain. "The sequence of sequencers: the history of sequencing DNA." Genomics 107.1 (2016): 1-8.

[17] Hira, Zena M., and Duncan F. Gillies. "A review of feature selection and feature extraction methods applied on microarray data." Advances in bioinformatics 2015 (2015).

[18] Kawaji, Hideya, Yosuke Yamaguchi, Hideo Matsuda, and Akihiro Hashimoto. "A graph-based clustering method for a large set of sequences using a graph partitioning algorithm." Genome Informatics 12 (2001): 93-102.

[19] Kchouk, Mehdi, and Faouzi Mhamdi. "New online hierarchical feature extraction algorithm for classification of protein." Database and Expert Systems Applications (DEXA), 2014 25th International Workshop on. IEEE, 2014.

[20] Kelil, Abdellali, Shengrui Wang, Ryszard Brzezinski, and Alain Fleury. "CLUSS: clustering of protein sequences based on a new similarity measure." BMC bioinformatics 8, no. 1 (2007): 286.

[21] Kingsford, Carl, and Rob Patro. "Reference-based compression of short-read sequences using path encoding." Bioinformatics 31, no. 12 (2015): 1920-1928.

[22] Korodi, Gergely, and IoanTabus. "An efficient normalized maximum likelihood algorithm for DNA sequence compression." ACM Transactions on Information Systems (TOIS) 23.1 (2005): 3-34.

[23] Li, Weizhong, and Adam Godzik. "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics 22.13 (2006): 1658-1659.

[24] Liu, Libin, Yee-kin Ho, and Stephen Yau. "Clustering DNA sequences by feature vectors." Molecular phylogenetics and evolution 41.1 (2006): 64-69.

[25] Mayilvaganan, M., and R. Rajamani. "Analysis of nucleotide sequence with normal and affected cancer liver cells using Hidden Markov model." Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on. IEEE, 2014.

[26] Montanari, Piero, Ilaria Bartolini, Paolo Ciaccia, Marco Patella, Stefano Ceri, and Marco Masseroli. "Pattern similarity search in genomic sequences." IEEE Transactions on Knowledge and Data Engineering 28, no. 11 (2016): 3053-3067.

[27] Nicolae, Marius, Sudipta Pathak, and Sanguthevar Rajasekaran. "LFQC: a lossless compression algorithm for FASTQ files." Bioinformatics 31.20 (2015): 3276-3281.

[28] Parsons, J. D., S. Brenner, and M. J. Bishop. "Clustering cDNA sequences." Bioinformatics 8.5 (1992): 461-466.

[29] Pettersson, Erik, Joakim Lundeberg, and Afshin Ahmadian. "Generations of sequencing technologies." Genomics 93.2 (2009): 105-111.

[30] Pinho, Armando J., Diogo Pratas, and Paulo JSG Ferreira. "Bacteria DNA sequence compression using a mixture of finite-context models." Statistical Signal Processing Workshop (SSP), 2011 IEEE. IEEE, 2011.

[31] Ramanujam, E., and S. Padmavathi. "Constraint frequent motif detection in sequence datasets." Advanced Computing (ICoAC), 2012 Fourth International Conference on. IEEE, 2012.

[32] Ren, Xianwen, et al. "iPcc: a novel feature extraction method for accurate disease class discovery and prediction." Nucleic acids research 41.14 (2013): e143-e143.

[33] Saha, Subrata, and SanguthevarRajasekaran. "Efficient algorithms for the compression of FASTQ files." In Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on, pp. 82-85. IEEE, 2014.

[34] Saidi, Rabie, Mondher Maddouri, and Engelbert Mephu Nguifo. "Protein sequences classification by means of feature extraction with substitution matrices." BMC bioinformatics 11.1 (2010): 175.

[35] Stojanov, Done, and Aleksandra Mileva. "A Short Survey of Pair-wise Sequence Alignment Algorithms." (2015): 237-242.

[36] Stranneheim, Henrik, Max Käller, Tobias Allander, Björn Andersson, Lars Arvestad, and Joakim Lundeberg. "Classification of DNA sequences using Bloom filters." Bioinformatics 26, no. 13 (2010): 1595-1600.

[37] Tembe, Waibhav, James Lowey, and Edward Suh. "G-SQZ: compact encoding of genomic sequence and quality data." Bioinformatics 26.17 (2010): 2192-2194.

[38] Wandelt, Sebastian, and Ulf Leser. "FRESCO: Referential compression of highly similar sequences." IEEE/ACM Transactions on Computational Biology and Bioinformatics10.5 (2013): 1275-1288.

[39] Wang, Jason Tsong-Li, Qicheng Ma, Dennis Shasha, and Cathy H. Wu. "New techniques for extracting features from protein sequences." IBM Systems Journal 40, no. 2 (2001): 426-441.

[40] Wendl, M. C., Korf, I., Chinwalla, A. T.,& Hillier, L. W. (2001). Automated processing of raw DNA sequence data. IEEE Engineering in Medicine and Biology Magazine, 20(4), 41-48.

[41] Yona, Golan, Nathan Linial, and Michal Linial. "ProtoMap: automatic classification of protein sequences and hierarchy of protein families." Nucleic acids research 28.1 (2000): 49-55.

[42] Yu, Qiang, Hongwei Huo, Xiaoyang Chen, Haitao Guo, Jeffrey Scott Vitter, and Jun Huan. "An efficient algorithm for discovering motifs in large DNA data sets." IEEE transactions on nanobioscience 14, no. 5 (2015): 535-544.

[43] Zhang, Zheng, Scott Schwartz, Lukas Wagner, and Webb Miller. "A greedy algorithm for aligning DNA sequences." Journal of Computational biology 7, no. 1-2 (2000): 203-214.

[44] Zhou, Hongxia, Liping Du, and Hong Yan. "Detection of tandem repeats in DNA sequences based on parametric spectral estimation." IEEE transactions on information technology in biomedicine 13.5 (2009): 747-755.

[45] Zhou, Qing, and Jun S. Liu. "Extracting sequence features to predict protein–DNA interactions: a comparative study." Nucleic acids research 36.12 (2008): 4137-4148.

Authors Profile

Ms.M.Muthu Lakshmi received Bachelor of Engineering from Anna University – Trichy in 2014 and Master of Engineering (Computer Science) from Manonmaniam Sundaranar University, Tirunelveli in 2016. She is currently pursuing Ph.D. in Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli. Her research work focuses on Bioinformatics and Computational Biology.

Dr.G.Murugeswari received Bachelor of Engineering (Computer Science) in 1995 and Master of Information Technology and Computer Engineering from Manonmaniam Sundaranar University, Tirunelveli. She completed her Doctorate in Computer Science and Information Technology in the year 2017. She is currently working as Assistant Professor in Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli. She has published 9 research papers in reputed international journals and 10 National and International conferences. Her area of interest includes Bio informatics and Computational Biology. Her area of specialization is Digital Image Processing.