International Journal of Latest Trends in Engineering and Technology
IJLTET Special Issue- ICRACSC-2016 , pp.067-071
e-ISSN: 2278-621X

USE OF DATA MINING TECHNIQUES AND STATISTICAL ANALYSIS FOR GENE EXPRESSION DATA AND INTERPRETATION

P. Mukesh[1] and Sameen Fathima[2]

Abstract: Use of data mining algorithms on sequence data can be used for charactering sequence. The popular mining tools like clustering and classification can be employed for analysis of EST data from various sources. Huge database for gene and EST sequence is available in public domain. The popular plant genomic resource databases are NCBI, Phytozome and Grameneetc.The classification functions perform differently across gene sequence. In machine learning, the classification is one of the popular data mining technique, considered as supervised learning, where training set of correctly identified observations are available.

Various analysestechniques like classification and sequential pattern miningwas carried out on gene sequence. There is scope to applying data mining algorithms/tools and statistical analysis. Grouping, clustering and characterizing sequence is important step in sequence analysis.Sequence is an enumerated collection of objects in which repetitions are allowed. Like a set, it contains members or also called elements, or terms. The number of elements, the same elements can appear multiple times at different positions in the sequence. The position of an element in a sequence is its rank or index, it is the integer from which the element is the image,it depends on the context or of a specific convention.

An expressed sequence tag (EST) is a stretch of DNA sequence which is used to recognize an expressed gene. The length of EST probable 200 to 1500 nucleotides or even more in a length, this count is sufficient to identify the full-length complementary DNA (cDNA). ESTs produced by sequencing a single segment of random clones from a cDNA library. The sequencing and analysis have allowed the rapid determination of many ESTs. Now, in sequence data bases, the majority of the sequences are ESTs.

Keywords – :Sequence analysis, protein, Web interface, data submission cDNA library,Eexpressed sequence tag (EST), normalization, Sequencing. Comparative genome, Classification, Single nucleotide polymorphism,Data mining technique.

I. INTRODUCTION

Sorghum bicolor (L.)Moenchis the cereal dry land crop. In sorghum crop cultivation and research it is required to improve sorghum yield and quality at huge amount of multiplication data is generated every year. As generated the data through by cropcultivation, seasonal experiments. Then it is required to compile and summarized the data which is used for present and future use for research work in sorghum crop. Data mining techniques selection have become an apparent need in many bioinformatics applications. In addition to the large pool of techniques that have already been developed in the machine learning and data mining fields, specific applications in bioinformatics have led to a wealth of newly proposed techniques.Data mining application success stories have been told in different areas among them healthcare, Banking and finance and telecommunication and in various subject sciences. Data mining is an emerging multidisciplinary field which facilitates discovering of previously unknown correlations, patterns and trends from large amounts of data stored in multiple data sources. It is a powerful new technology with great potential to help businesses and science. Several major data mining techniques have been developed those methods are association, classification and clustering, prediction and sequential patterns etc.

Plant genomics projects involving model species and many agriculturally important crops are resulting in a rapidly increasing database of genomic and expressed DNA sequences.The publicly available collection of

[1] *Indian Institute of Millets Research, Rajendranagar, Hyderabad, TS, mobile 9440956548*
[2] *College of Engineering, Osmania University, Hyderabad, TS*

expressed sequence tags (ESTs) from several grass species can be used in the analysis of both structural and functional relationships in these genomes. Functional analysis may reveal their role in plant metabolism and gene evolution.
Genes, the human gene compendium, enables researchers to effectively navigate and inter-relate the wide universe of human and plant genes, diseases, variants, proteins, cells, and biological pathways. There are many sequence alignment software are available for sequence analysis. The widely used algorithm for alignment is CLUSTAL-W which allows multiple sequence alignment. The Expressed sequence tag (EST) of genes available in public domain can be surveys and can be efficiently characterized. The data mining algorithms can be developed to study the sequence homology, allelic variation, repetitive sequence characterization etc.But the available sequence alignment software is a compilation of software tools and web portals used in pair wise sequence alignment and multiple sequence alignment and structural alignment for structural alignment of proteins. Similarly, a program can be developed to find simple sequence repeats (SSR) marker in each gene/genome and identify single nucleotide polymorphism (SNP) in expressed sequence tag sequences. Data sources are available from open source web environment. That is specifically from NCBI or any other source will be collected necessary input data from specified sources.

II. PROPOSED ALGORITHM

Data mining application success stories have been told in different areas among them healthcare, Banking and finance and telecommunication and in various subject sciences. Data mining is an emerging multidisciplinary field which facilitates discovering of previously unknown correlations, patterns and trends from large amounts of data stored in multiple data sources. It is a powerful new technology with great potential to help businesses and science.
Data Mining is the process of sifting through stores of data to extract previously unknown, valid patterns and relationships that provide useful information. Data Mining uses sophisticated data analysis tools and visualization techniques to segment the data and evaluate the probability of future events.

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classification normally uses prediction rules to express knowledge expressed in the form of IF-THEN rules, where the antecedent (IF part) consists of a conjunction of conditions and the rule consequent (THEN part) predicts a certain predictions attribute value for an item that satisfies the antecedent. Types of Classification Models are Neural Networks, Support Vector Machines, Bayesian classifiers etc. On the contrary, decision trees and rule classifiers have a similar operational profile. Classification methods are typically strong in modeling interactions. Several of the classification methods produce a set of interacting loci that best predict the phenotype.

The association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, RakeshAgrawal et al, a typical and widely-used example of association rule mining is Market Basket Analysis. The problem is to generate all association rules that have support and confidence greater than the user-specified minimum support and minimum confidence. Association rule mining algorithms include; Apriori, AprioriTid, Apriori hybrid and tertius algorithms. Sequential Pattern Mining Sequential pattern mining deals with finding statistically relevant patterns between data examples where the values are delivered in sequence. It is closely related to time series mining and special case of structural data mining. Some of the applications are analysis of customer purchase patterns or Web access patterns, analysis of time related processes involved in scientific experiments, disease treatments, DNA sequencing etc. Classification of sequential pattern mining algorithms.

The three main categories of Sequential pattern mining algorithms that have been in use are as follows Apriori based algorithms such as GSP, SPADE, SPAM algorithms, pattern growth algorithms such as Free Span and Prefix Span early pruning algorithms such as LAPIN-SPAM.

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.
Gene supplies gene-specific connections in the nexus of map, sequence, expression, structure, function, citation, and homology data. Unique identifiers are assigned to genes with defining sequences, genes with known map positions, and genes inferred from phenotypic information. These gene identifiers are used throughout databases and tracked through updates of annotation. Gene includes genomes represented by Sequences are integrated for indexing and query and retrieval from database systems.

Gene is accessed like any database, namely by, querying on any word, restricting the query term to a certain field and applying filters or properties. The following are sequence input database formats.

EST SEQUENCE of Sorghum crop

GCATCCCCGAGTTCTTGTGAGGGACAAGCGTGGAGAGTTTATGAAGGGAAGACCTCCGGTATTTACACACTCAGCTGATCCCATGGAGGCAGATGAT
TGGTTACGTGCTGTGGAGAGGCAGCTAAACATAGCACAGTGCAATGACTTGGAGAAGGTGTTGTATGCTTCTGGACAGCTTCAAGGTGCAGCTCAGA
CATGGTGGGAGTCATATCAAGCTGCTCGTCCCAACAAATGATCCTCCTATCACATAGCTGGAATTCTATAGGGACTTCAGAGCTCGACACATGAACT
CGGGGGATGAC

I. EST02397 FASTA FORMAT FROM EXPRESS HEART CDNA LIBRARY CDNA CLONE, MRNA SEQUENCE, SORGHUM GENE INSTANCE.>CV429168.1 EST02397 ATLANTIC SALMON LAMBDA ZAP EXPRESS HEART CDNA LIBRARY SALMOSALARCDNA CLONE HR_008_H02 5', MRNA SEQUENCE

GGCACGAGGATTACACACAGGTGTGTTAGTCAAATATGTTGCCTGTTAGATTAGAACTGCTGAAGTTATTTTCTGAAGGGATCGTTTTATATGAAGT
AACTAGATTAAAAAAAAAAGGAAAAAAAAGATGGTAAAAGGGTGAGTTCTGTCAATTGCTATTGATTTTCACACTTGCACATAGTTTAGAACTTCA
CATACTATAGATTTATATTTGTCTAAGATGAAGCACTTTGTCCCTGGTCAACTCCTTTAAGAATAAAATGTATTTTGTTTGTCTTCTGAAAAAAAAA
AAAAAAAAAA

SSR survey of rice, maize, poplar, tomato, cotton, and soybean EST sequences
Numbers in parentheses show percentage of total SSR content.

SSR survey of rice, maize, poplar, tomato, cotton, and soybean EST sequences

Source	Rice	Maize	Soybean	Tomato	Cotton	Poplar
Di	657 (13)	140 (18)	147 (30)	84 (21)	53 (22)	38 (28)
Tri	3,747 (73)	478 (61)	311 (63)	289 (72)	157 (66)	83 (61)
Tetra	498 (10)	126 (16)	30 (6)	24 (6)	21 (9)	14 (10)
Penta	230 (4)	46 (6)	9 (2)	2 (1)	8 (3)	1 (1)
Total SSR	5,132	790	497	399	239	136
No.	45,033	14,950	9,611	9,100	8,083	4,809

Source	Rice	Maize	Soybean	Tomato	Cotton	Poplar
sequence						
Average length	380	430	380	490	590	390
Total length (kb)	17,304	6,411	3,675	4,444	4,788	1,880
Average distance (kb)	3.4	8.1	7.4	11.1	20.0	14.0

Numbers in parentheses show percentage of total SSR content.

The sequencing of RNA also has transitioned and now includes full-length cDNA analyses, serial analysis of gene expression (SAGE)-based methods, and noncoding RNA discovery. Next-generation sequencing has also enabled novel applications such as the sequencing of ancient DNA samples, and has substantially widened the scope of metagenomic analysis of environmentally derived samples. Taken together, an astounding potential exists for these technologies to bring enormous change in genetic and biological research and to enhance our fundamental biological knowledge.

Needed to develop a computer based method to identify candidate single nucleotide polymorphisms (SNPs) and small insertions/deletions from expressed sequence tag data. Using a redundancy-based approach, valid SNPs are distinguished from erroneous sequence by their representation multiple times in an alignment of sequence reads. A second measure of validity was also calculated based on the cosegregation of the SNP pattern between multiple SNP loci in an alignment. After analysis the candidate polymorphisms were identified with an SNP redundancy score of two or greater. Segregation of these SNPs with haplotype indicates that candidate SNPs with high redundancy and cosegregation confidence scores are likely to represent true SNPs. The SNP transition/transversion ratio and insertion/deletion size frequencies correspond to those observed by direct sequencing methods of SNP discovery and suggest that the majority of predicted SNPs and insertion/deletions identified using this approach represent true genetic variation in maize.

III. EXPERIMENT AND RESULT

IV.CONCLUSION

The advent and widespread availability of next-generation sequencing instruments has ushered in an era in which DNA sequencing will become a more universal readout for an increasingly wide variety of front-end assays. However, more applications of next-generation sequencing, beyond those covered here, are yet to come. For example, genome resequencing will likely be used to characterize strains or isolates relative to high-quality reference genomes such as *C. elegans*, *Drosophila*, and human.

Studies of this type will identify and catalog genomic variation on a wide scale, from single nucleotide polymorphisms (SNPs) to copy number variations in large sequence blocks (>1000 bases). Ultimately, resequencing studies will help to better characterize. Hence the mining technique like associate sequential mining analysis is widely used in plant genomic sequences better than other organisms.

REFERENCES

[1]. Victor L. Jong1, Putri W. Novianti, Kit C.B. Roes1 and Marinus J.C. Eijkemans1, Date: 11 February 2016, "selecting a classification function for class prediction with gene expression data", Bioinformatics Advance Access,published,February 18, 2016.

[2]. CristianTaccioli, VincenzaMaselli, JesperTegne, David Gomez-Cabrero, GioiaAltobelli, Warren Emmett, Francesco Lescai, Stefano Gustincich and Elia Stupka1, Database, Vol. 2011, Article ID bar007, doi:10.1093/database/bar007, "ParkDB: a Parkinson's disease gene expression, database".

[3]. Carol J. Bult*, Janan T. Eppig, Judith A. Blake, James A. Kadin, Joel E. Richardson and theMouse Genome Database Group *Vol. 44, Database issue Published online 17 November 2015,* "Mouse genome database", 2016*D840–D847 Nucleic Acids Research, 2016, doi: 10.1093/nar/gkv1211.*

[4]. Sumudu N Dissanayake, OsvaldoMarinotti, Jose Marcos C Ribeiro and Anthony A James, Received:12 January 2006, Accepted: 17 May 20063, Published:17 May 2006, "angaGEDUCI: *Anopheles gambiae*gene expression database with integrated comparative algorithms for identifying conservedDNAmotifs in promoter sequences".

[5]. RajuBhukya, DVLN Somayajulu, "Exact Multiple Pattern Matching Algorithm using DNA Sequence and Pattern Pair", *International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March. 201.*

[6]. M. Blatt , S. Wiseman , and E. Domany , "Super-Paramagnetic Clustering of Data," *Physical Rev. Letters,* vol. 76, 1996.

[7]. Rohlf FJ (2000) NTSYS-pc, numerical taxonomy and multivariate analysis system, version 2.10e. Applied Biostatistics, New York.

[8]. J.B. MacQueen , "Some Methods for Classification and Analysis of Multivariate Observations," *Proc. FifthBerkeley Symp.Math. Statistics and Probability,* pp. 281-297, Univ. of California, Berkeley, Univ. of California Press, Berkeley, 1967.

[9]. Data1 Jacqueline Batley2 etal ., Agriculture Victoria, Plant Biotechnology Centre, La Trobe University, Bundoora, Victoria 3086, Australia (J.B., D.E.); and the School of Biological Sciences, University of Bristol, Bristol B58 1UG, United Kingdom (G.B.,H.O.,K.J.E.), "Mining for Single Nucleotide Polymorphisms and Insertions/Deletions in Maize Expressed Sequence Tag"

[10]. Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol48:501–510,.

[11]. Boguski MS, Lowe TM, Tolstoshev CM: dbEST - database for expressed sequence tags! NafGenef 1993,4:332-333.

[12]. Cooke R, Raynal M, Laudie M, Grellet F, Delseny M, Morris PC, Guerrier D, Giraudat J, Quigley F, Clabault G et a/.: Further progress towards a catalogue of all Arabidopsis genes: analysis of a set of 5000 non-redundant ESTs. Plant 11996, 9:101-l 24.

[13]. Parh DK, Jordan DR, Aitken EAB, Mace ES, Jun-ai P, McIntyre CL, Godwin ID (2008) QTL analysis of ergot resistance in sorghum. TheorAppl Genet 117:369–382 Altschul,S. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402.

[14]. Orland Gonzalez1,2,* and Ralf Zimmer, April 1, 2008, Assigning functional linkages to proteins using phylogeneticprofiles and continuous phenotypes.

[15]. **R. Agarwal, et.al. "Mining sequential patthers,** Data Engineering, 1995. Proceedings of the Eleventh International Conference on, ,**IEEE *Xplore*:** 06 August 2002, **ISBN Information: INSPEC Accession Number:** 4886436,