Journal Homepage: -www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI:10.21474/IJAR01/7270
DOI URL: http://dx.doi.org/10.21474/IJAR01/7270

RESEARCH ARTICLE

COMPUTATIONAL ANALYSIS ON GENE EXPRESSION PATTERN: A SURVEY.

K. Vimala[1] and Dr. D. Usha[2].
1. Research Scholar Department of Computer Science Mother Teresa Women's University.
2. Assistant Professor Department of Computer Science Mother Teresa Women's University.

Manuscript Info

Manuscript History

Received: 12 April 2018
Final Accepted: 14 May 2018
Published: June 2018

Keywords:-
Gene, DNA, Microarray, Protein structure, Gene expression, Computational Analysis.

Abstract

A gene is a part of Deoxyribonucleic acid (DNA), which contains all the information required to analysis the defects and genetic problems that evolves in an organism. A gene is also the element of information that is transferred through transcription and translation. This paper discusses the transformation that happens in the cell either internal or external environment, that can lead to changes in gene expression.Most human and animal diseases are manifested through a mis-regulation of gene expression. The outputs of DNA Microarray are processed by computation tools to take out biological significance which may help to detect human disease. Computation tools include a variety of algorithms of data mining, pattern recognition and machine learning etc. Finding desired algorithm plays a major role in research to satisfy the requirements. Computational analysis on supervised, unsupervised and semi supervised classification are considered for survey.

Introduction:-

Disease gene identification is a process of identifying the mutant genotypes that are responsible for an inherited genetic disorder. Mutations in these genes can include single nucleotide substitutions, single nucleotide additions/deletions, deletion of the entire gene, and other genetic abnormalities. A disease gene identification technique follows few procedures. DNA is first collected from several patients who are believed to have the same genetic disease. Then, their DNA samples are analyzed and screened to determine probable regions where the mutation could potentially reside. These probable regions are then lined-up with one another and the overlapping region should contain the mutant gene. If enough of the genome sequence is known, that region is searched for candidategenes. Coding regions of these genes are then sequenced until a mutation is discovered or another patient is discovered, in which case the analysis can be repeated, potentially narrowing down the region of interest.

Gene Disorder:-

A genetic disorder is a genetic problem caused by one or more abnormalities in the genome. Most genetic disorders are quite rare and affect one person in every several thousands or millions. Genetic disorders may be hereditary, passed down from the parents' genes. In other genetic disorders, defects may be caused by new mutations or changes to the DNA. In such cases, the defect will only be passed down if it occurs in the germ line.

Corresponding Author:-K. Vimala.
Address:-Research Scholar Department of Computer Science Mother Teresa Women's University.

Hereditary Disorders:-
Single gene inheritance, also called Mendelian or monogenetic inheritance. This type of inheritance is caused by changes or mutations that occur in the DNA sequence of a single gene. Multifactorial inheritance, which is also called complex or polygenic inheritance. Multifactorial inheritance disorders are caused by a combination of environmental factors and mutations in multiple genes. Multifactorial inheritance, which is also called complex or polygenic inheritance. Multifactorial inheritance disorders are caused by a combination of environmental factors and mutations in multiple genes.

Gene Expression analysis:-

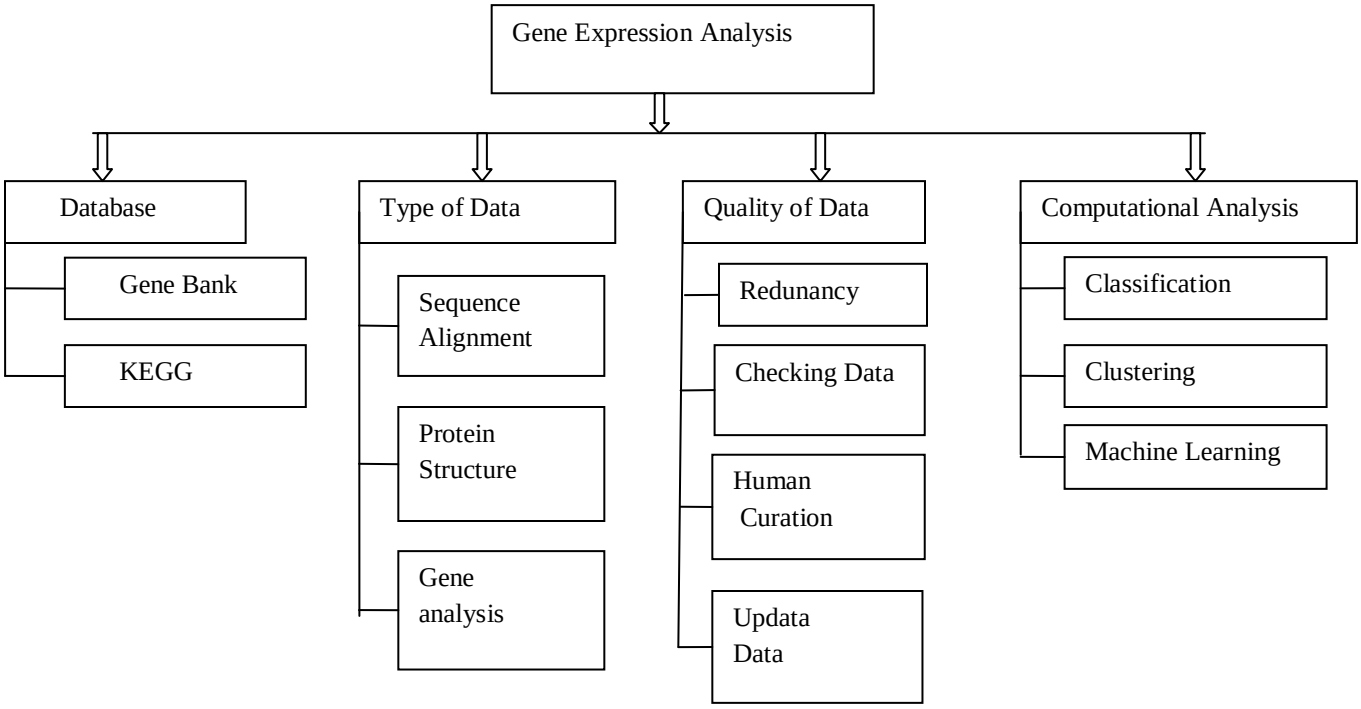

Fig:- Gene Expression Analysis

Database:-
These gene databases are very large in size and complex to work with, hence to store, access and manipulate these data efficiently is important deal. Gene databases are categorized as sequence databases, genome databases, microarray databases, protein structure databases and many more. The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information.

KEGGgenes are a collection of gene catalogs for all complete genomes generated from publicly available resources. The database contents [11] represent two main challenges a) Hierarchies of Co-expressed Genes and Coherent Patterns. b) Address the High Connectivity of Gene Expression Data Sets.

Bioinformatics:-
Sequence analysis and alignment:-
The most well-known application of bioinformatics is sequence analysis. In sequence analysis, DNA sequences of various organisms are stored in databases for easy retrieval and comparison. DNA sequences used for bioinformatics can be collected in a number of ways. One method is to go through a genome and search out individual sequences to record and store. Another method is to compare all fragments for finding whole sequences by overlapping the redundant segments. Sequence alignment [11], is an element method of information management, it has all important sense for discovering biologic sequence function, structure and evolution. Sequence [10] subsets are identified using bisecting-kmeans algorithm where K-mer counts are considered as attributes for clustering.
Prediction of protein structure:-

Proteins play crucial functional roles in all biological processes: enzymatic catalysis, signaling messengers, structural elements. Function depends on unique 3-D structure. It is easy to obtain protein sequences but difficult to determine structure. Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. Structural information of protein structure is usually classified as one of secondary, tertiary and quaternary structure. Protein structure prediction is the prediction of the three-dimensional structure of a protein from its amino acid sequence i.e, the prediction of its tertiary structure from its primary structure. Protein structures are being determined with increasing speed. Protein Structure Prediction [13] is the process of predicting the three dimensional structure of a protein from its amino acid sequence. A layered architecture [14] with two interacting levels has been defined for dealing with both primary and secondary-structure information of target protein sequences.

Gene expression:-

Genes encode proteins and proteins dictate cell function. Moreover, each step in the flow of information from DNA to RNA to protein provides the cell with a potential control point for self-regulating its functions by adjusting the amount and type of proteins it manufactures. A novel chromosome representation [5] involves each chromosome to be embedded with sub-functions, which can be deployed to construct the final solution. As part of the chromosome, the sub-functions are self-learned or self-evolved. Genes [17] have several distinctive roles in cellular processes; this is very difficult problem for classical clustering methods so mixture model is used to avoid this problem, with hidden Markov models (HMMs) as effective and flexible components

Microarray Analysis Of Gene Expression:-

Gene analysis:-

Genetic analysis refers to identification of genes influencing physical characteristics in organisms and their patterns of inheritance. Major research is concentrated for analysis of this interpreted data. Number of tools and techniques are available for analysis purpose like DNA Microarray, SAGE, Tiling array etc.

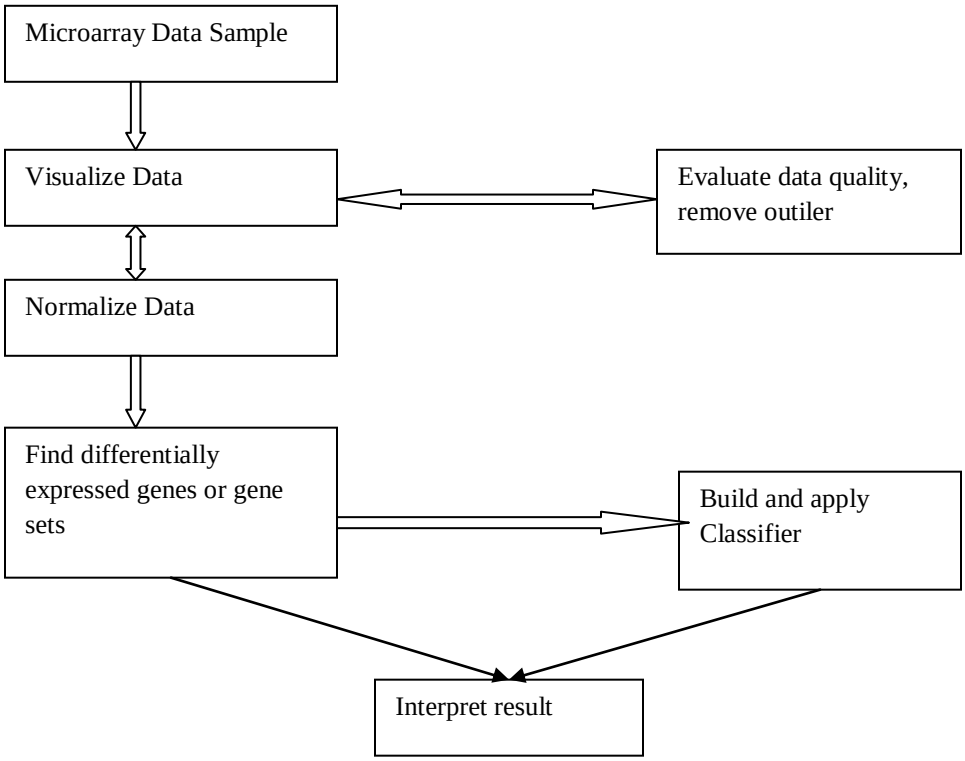

Fig :- Microarray Analysis

Microarray:-

Microarray provides a foundation to genotype, thousands of different loci at a time, which is useful for association and linkage studies to isolate chromosomal regions related to a particular disease. This array can also be useful to locate chromosomal abnormalities related to disease. The different DNA fragments are arranged in rows and columns such that the identity of each fragment is known through its location on the array. Two types of microarrays are gene expression microarray and tissue microarray.

Microarray analysis of gene expression:-

Microarray analysis allows description of genome-wide expression changes in health and disease. Microarrays can be broadly classified according to at least three criteria: 1) length of the probes; 2) manufacturing method; and 3) number of samples that can be simultaneously profiled on one array. Microarray technologies [1] have provided the means to monitor the expression levels of a large number of genes simultaneously. Gene clustering and gene ordering are important in analyzing a large body of microarray expression data.

Computational Analysis:-

Computational tools:-

The scale and complexity of genetic and genomic data are ever-expanding, requiring biologists to apply increasingly more sophisticated computational tools in the analysis, interpretation and storage of these data. Data mining, Machine learning and Pattern recognition are some of the computational tools required for to analysis the database. Data mining extracts needed data from a larger set of any raw data. It implies analyzing data patterns in large batches of data using one or more software. The analyses are undergone with the help of clusters analysis and frequent pattern analysis to find hidden pattern in data samples [19]. The patterns identified in the data suggest similarities in the gene behavior, which provides useful information for the gene functionalities. Patterns are focused on two important methods, Supervised and Unsupervised learning [6]. While machine learning is an artificial intelligence part which focuses on complex pattern using statistics, probability theory, or artificial intelligence. The decisions are made on this identified data.

Computational Analysis of Gene Expression:-

Gene computational analyses are required categories the gene based on their behavior. Classifications of genes are given as, supervised, unsupervised and semi supervised methods. Supervised method includes all data labeled and the algorithms learn to predict the output from the input data. Unsupervised method includes all data unlabeled and the algorithms learn to inherent structure from the input data. Semi-supervised includes some data labeled but most of it is unlabeled and a mixture of supervised and unsupervised techniques can be used. K-means clustering algorithm, run fast and consume less memory compared to hierarchical clustering algorithms. [20] Poisson-based measures and K-means clustering algorithm is to group tags with similar count profiles across libraries. An effective ensemble [21] approach is proposed. Ensemble classifier increases the performance of the classification, and also improves the confidence of the results. The ensemble classifiers results are very less dependent on peculiarities of a single training set. Semi-supervised classification [13] improved to be the best prediction accuracy method. Support Vector Method performance increased with the number of unlabeled samples.

Conculsion:-

An efficient system can be developed to handle gene expression changes based the disorders related to heart disease such as blood pressure, congenital heart disease and cardiac attack and diabetic for animals and human. The system involves two-level analysis. The first level includes preprocessing techniques, where the redundancies in the dataset are removed and summarizations of the data are collected. The second level includes statistical analysis, classification and development of ontology. The development in gene expression can provide an opportunity to study features of disorder in disease and provide path physiological context to handle complex disease.

Gene Expression Analysis:-

Table:- Summary for Gene Expression Analysis

Summary for Gene Expression Analysis			
S.no	**Evaluation**	**Ref**	**Objective**
1.	Sequence analysis and alignment	11	Sequence alignment, is an element method of information management, it has all important sense for discovering biologic sequence function, structure and evolution.
		10	Sequence subsets are identified using bisecting-kmeans algorithm where K-mer counts are considered as attributes for clustering.
2.	Prediction of protein structure	13	Protein Structure Prediction is the process of predicting the three dimensional structure of a protein from its amino acid sequence.
		14	A layered architecture with two interacting levels has been defined for dealing with both primary and secondary-structure information of target protein sequences.
3.	Gene expression	5	A novel chromosome representation involves each chromosome to be embedded with sub-functions, which can be deployed to construct the final solution. As part of the chromosome, the sub-functions are self-learned or self-evolved.
		17	Genes have several distinctive roles in cellular processes; this is very difficult problem for classical clustering methods so mixture model is used to avoid this problem, with hidden Markov models (HMMs) as effective and flexible components
4.	Computational Analysis	20	Poisson-based measures and K-means clustering algorithm is to group tags with similar count profiles across libraries.
		21	An effective ensemble approach is proposed. Ensemble classifier increases the performance of the classification, and also improves the confidence of the results. The ensemble classifiers results are very less dependent on peculiarities of a single training set.
		13	Semi-supervised classification improved to be the best prediction accuracy method. Support Vector Method performance increased with the number of unlabeled samples

Reference:-

1. Yuan-Fang Tsai, Jinn-Moon Yang, Huai-Kuang Tsai, and Cheng-Yan Kao,(2004) “An Evolutionary Approach for Gene Expression Patterns”, IEEE Transactions On Information Technology In Biomedicine, Vol.8, No.2, pp.69.
2. van Houwelingen HC, de Menezes RX, Boer JM,(2009) “Microarray Data Analysis”. Applied Bioinformatics. Vol 3, Issue 4, pp. 229.
3. Yew-Soon Ong, Wentong Cai and Jinghui Zhong,(2015), ‘Self-Learning Gene Expression Programming”, IEEE Transactions On Evolutionary Computation, Vol.2, No.3,Pages.23, 2015
4. Wang P-C, Su C-T, Chen L-F, Chen K-H.(2012) “Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis”. Neural Computing and Application, Vol 21, Issues 8 , pp. 2087.
5. Shuanhu Wu, Alan Wee-Chung Liew, (2004)“Cluster Analysis of Gene Expression Data Based on Self-Splitting and Merging Competitive Learning”, IEEE Transactions On Information Technology In Biomedicine, Vol. 8, No. 1, pp.234.
6. Safiye Celik, Scott M. Lundberg, Benjamin A. Logsdon,(2018), “A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia”,Nature CommunicationsVol. 9, Article number: 42.
7. Xiaojie Lu, Mingquan Ye , Lingyun Gao , Daobin Huang,(2017) “Hybrid Method Based on Information Gain and Support Vector Machine for Gene Selection in Cancer Classification”,Genomics Proteomics Bioinformatics, Vol 15, No.1, pp. 389.
8. Wang R , Miao D ,Chen Y,(2010) “A rough set approach to feature selection based on ant colony optimization”. Pattern Recogniton Letters, Vol 31, No.3, pp. 226.

9. Thusangi Wannige, Kokila K. Perera, C. (2016)"A Hybrid Algorithm for Multiple DNA Sequence Alignment", International Conference on Advances in ICT for Emerging Regions, Vol.3, No.5,pp. 323.
10. LIU Shuai, LIU Chao,(2011) "The research on DNA Multiple Sequence Alignment Based on Adaptive Immune Genetic algorithm", International Conference on Electronics and Optoelectronics.
11. Jian Pei, Daxin Jiang, (2005), "An Interactive Approach To Mining Gene Expression Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 10, pp.456.
12. Sankar K. Pal, "Evolutionary Computation in Bioinformatics(2006): A Review", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 36, No. 5, pp. 601.
13. Mingguang Shi and Bing Zhang,(2011)"Semi-supervised learning improves gene expression-based prediction of cancer recurrence", BioinformaticsVol. 27, No. 21, pp. 3017.
14. Subhendu Bhusan Rout, Satchidananda Dehury, Bhabani Sankar Prasad Mishra,(2013) "Protein Structure Prediction using Genetic Algorithm", International Journal of Computer Science and Mobile Computing, Vol.2, Issue.6, pp. 187.
15. Luciano Milanesi, Giuliano Armano, and Alessandro Orro, (2005),"Multiple Alignment Through Protein Secondary-Structure Information", IEEE Transactions On Nanobioscience, Vol.4, No.3, pp.34.
16. Osowski S, Latkowski T.(2015) "Data mining for feature selection in gene expression autism data." Expert Systems with Applications, Vol 42, Issues 2, pp. 864.
17. Sushmita Mitra,(2012)"Feature Selection and Clustering of Gene Expression Profiles Using Biological Knowledge", IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 42, No. 6, pp.1590.
18. , Ivan G. Costa, Christine Steinhoff, Alexander Schliep and Alexander Schonhuth,(2005) "Analyzing Gene Expression Time-Courses", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 2, No. 3, pp. 179.
19. Francisco Torres Aviles,(2014)." Data mining and influential analysis of gene expression data for plant resistance gene identification in tomato (Solanum lycopersicum)",Electronic Journal of BiotechnologyVol 17, Issue 2, pp. 79.
20. Wong, W. H , S, Liu, Cai, L, Huang, H, Blackshaw, J. S, Cepko, C. L.(2004) "Clustering Analysis of SAGE Data Using a Poisson Approach," Genome Biology, Vol.5, No.51, pp.56.
21. Sara Tarek , Mahmoud Shoman, Reda Abd Elwahab, (2016)"Gene expression based cancer classification", Egyptian Informatics Journal, Vol.4, No.3, pp.234.
22. David Seo, MD, Geoffrey S. Ginsburg, (2006)"Gene Expression Analysis of Cardiovascular Diseases Novel Insights Into Biology and Clinical Applications", Cardiovascular Genomic Medicine, Vol. 48, No. 2, pp. 227.
23. Michelle M. Kittleson, Khalid M. Minhas,(2005) "Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure",Physiol Genomics, Vol. 21, No.4, pp. 299.
24. Yuting Liu, Marko Briesemann, Jun Yan, Haifang Wang,(2010) , "Computational analysis of gene regulation in animal sleep deprivation",Physiol Genomics,Vol.42,No.2,pp. 427.