ISSN: 2393-8528

IJICSE
Open Access
Journal Approved by UGC

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 4 Issue 4; July-August-2017; Page No. 29-35

A Study on Human Genome Sequences for Detecting Inherited Diseases Using Optimized Methods

J.Viba Mary

Assistant Professor, Dept of M.Sc. (SS & SS), KG College of Arts & Science, Coimbatore-641035.

vibamphil@gmail.com

Received 10 May 2017; Accepted 04 July. 2017

ABSTRACT

This research is concerned with the study of the gene and analysis of the Human Genome using Sequence Analysis efficiently and effectively in order to detect the inherited diseases carrying gene. The Genome sequencing, especially in humans, has been helpful in identifying inherited diseases in human body. An optimized method is used to predict these inherited disease causing genes located in DNA. This paper describes the existing methods developed for detecting the inherited diseases in the gene. The objective of this paper is to summarize and compare the well known methods applied in finding the inherited Diseases has made become a valuable one in medical field as well as to the society.

Keywords: Human Genome Sequences, Inherited diseases, Optimized methods, Sequence Analysis

I. INTRODUCTION

Human Genome Sequences have been used for identifying genes causing diseases located in the DNA of human body [1]. Inherited diseases have been affecting human beings to a greater extent and are many procedures in the biological methods have been in practice for identifying inherited disease causing genes [24]. A technique which allows the researchers to collect information about the genes found in the DNA is called Genomic sequencing [21].

Genomic sequencing is done in DNA of all plants, animals, bacteria and in the humans. This technique which used for sequencing a genome especially in humans is called as Human Genome Sequencing [21]. Adenine, Guanine, Cytosine and thymine is referred by letters A, G, C and T found in DNA [22]. Each and every nucleotide is made up of nitrogen based material such as guanine (G), adenine (A), thymine (T), and cytosine (C). Many techniques are available for sequencing these letters such that G, A, T, and C in the genes [23].
There are various techniques to read the sequences of these letters in genes and genomes. The sequenced genes of human are represented in the fig.1 by the letters A, G, T and C.

1. The below figure is depicted from [4].

Fig.1: Genome Sequencing

Genome Sequencing is also called as decoding, but a sequence is still very much in code. Hence, a genome sequence is said to be a very long string of letters in a mysterious language. Researchers are still working for the translation of those strings of letters in the way such that how the genome is working , what are all the various genes that make up the genome to do, how different genes are related, and how the various parts of the genome are coordinated [4].
The database which provides the information relevant to the inheritary Diseases can be obtained from NCBI (http://www.ncbi.nlm.nih.gov /Entrez), PubMed, GenBank, EMBL (http:// www.ebi.ac.uk), and DDBJ (http://www.ddbj. nig.ac.jp) [16]. There are many sequencing methods are available in biological methods for sequencing genes. But here, along with the biological methods, Computer Assisted Methods which is used for identifying inherited diseases are discussed. Then a comparison is made between those methods in order to make the research in a better way.

*Corresponding author: J.Viba Mary

This paper is organized as follows: Section II describes the methods used for identifying the inherited diseases. Section III illustrates the comparison between those methods to make the researchers in a right way to lead their research. The final part of the paper presents the conclusions.

II. METHODOLOGY

T Wei Huang et al. (2014) proposed a possible cure of dimensionality. Radial Bias Neural Network and the fuzzy clustering is employed which provided a higher degree of accuracy when compared to the other models which provided accuracy earlier [2]. Zolton Tezso et al. (2009) employed a novel computational methodology to find the interactions between protein–protein. The predicted regulatory nodes are used to find a more effective way to reconstruct the disease pathways. The results showed combinational treatment strategies for a wide range of diseases.
Yili Liu et al. (2014) developed a new clustering algorithm for cancer classification namely the Network Assisted Co-clustering to identify cancer sub-types [3]. This study has proved that this tool is useful for identifying cancer subtypes with a high through-put and the high dimensional gene expression data. Usha kuppusamy et al. (2014) carried out Baseyian inference in Gene Ontology by employing Graphic structure and the protein – protein interactions [6]. This approach is being used to predict the cellular component and the molecular functions.

Peter et al. (2008) introduced Human Phenotype Ontology in which represents the Phenotypic Anomalies and it was effective in capturing the phenotypic similarities between the diseases in high efficient manner [6]. Anuradha et al. (2014) proposed Factual Dimension Method along with the Practical Swarm Optimization (PSO) for clustering the samples into groups and also for selecting minimum threshold value.

Yuhai Zhao et al (2014) identifies that the problem in microarray data analysis is to discover the phenotype structures. The goal was to find the samples corresponding to different phenotypes (such as disease or normal), and for each group of samples [7]. Following this, the representative expression pattern or signature that distinguishes this group from others was also to be found. An efficient algorithm, FINDER was developed to improve the accuracy of phenotype structures discovered and detects signatures with a high discriminative power. Xiujuan Wang et al (2011) applied the molecular network concept to identify the similarities encountered in the phenotypic structure [7].

Timothy et al. (2014) used a top-down approach to determine the rates of loss or gain of known human exonic splicing regulatory (ESR) sequences associated with either disease-causing mutations or putatively neutral single nucleotide polymorphisms (SNPs) [10]. It was discovered that using a top-down approach to determine rates of loss or gain of known human exonic splicing regulatory (ESR) sequences associated with either disease-causing mutations or putatively neutral single nucleotide polymorphisms (SNPs). It was discovered further, around ~25% (7154/27,681) of known mis-sense and nonsense disease-causing mutations were found to alter functional splicing signals within exons, suggesting a much more widespread role for aberrant mRNA processing in causing human inherited disease than that has hitherto been appreciated [9,10].

Cho-Jui Hsieh et al. (2014) proposed a Divide and Conquer Kernal Solver Support Vector Machine to offer using classification methods that what was reported in. SVM is one of the best suited for Classification Methods [11]. The study proved that the DC-SVM was capable of providing a higher degree of accuracy when compared to others. Quan Zhong et al (2009) found that a single gene, to multiple disorders, can be modelled by distinguishing edgetic network perturbations [12]. Edgetic network perturbation models might improve both the understanding of dissemination of disease alleles in human populations and the development of molecular therapeutic strategies.
Yong Lu et al (2008) developed a generative probabilistic model which identifies a subset of categories that explained the selected gene set. The current model accommodates noise and errors in the selected gene set and GO [13]. Using controlled GO, this method recovered data from most of the selected categories accurately, leading to dramatic improvement when compared to the current methods for GO analysis. When microarray expression data and chip-chip data are used from yeast and human, it helps in identifying both general and specific enriched categories [13].

Igor Feldman et al (2007) shows that the disease mutations are less likely to occur in essential genes when compared to other human gene and the diseased genes display a significant functional clustering in the analyzed molecular network [14].

Thus, Functional Clustering proves to be a more effective and improved method to identify yet-to-be discovered genes.

A. BIOLOGICAL METHODS

The biological methods that have been used in this study for identifying the inherited diseases causing gene in the human body have been described in the subsequent paragraphs.

a) Pre-genomics Techniques

Pre-genomics techniques are focused at the selective regions of the genome with minimum knowledge about the gene sequences [8]. Genetic techniques are those techniques which are capable of providing information related regarding Restriction Fragment Length Polymorphism (RFLP) and microsatellite analysis.

b) Loss of Heterozygosity (LOH)

Loss of Heterozygosity (LOH) is a technique used for comparing two samples from the same patient. LOH analysis is also used under the condition which that cancer causing genes is identified in one sample called mutant sample consists of tumor DNA and the other sample is a control sample which consists of genomic DNA from non-cancerous cells of the same patient. RFLPs and microsatellite markers provides DNA polymorphism patterns which is then interpreted in a heterozygous region or a homozygous region of the genome [9]. The deletion of a single copy of the same gene resulted that all patients are affected with the same disease. All patients contain two regions out of which one is a heterozygous region composed of control sample and the other is a homozygous region composed of mutant sample [5]. It can be concluded that the homozygous region containing the mutant sample always has the diseased gene.

c) Post-Genomics Techniques

High-Throughput Sequencing and software capable of genome-wide analysis is the latest sequencing technology which have made sequences cost effective and less time-consuming [8]. It also provides more efficient identification techniques to identify the disease- causing genes.

d) Identity by Descent Mapping

Identity by Descent (IBD) Mapping is most commonly used as a single nucleotide polymorphism (SNP) arrays to survey the known polymorphic sites throughout the genome of affected patients. When these SNPs do not cause any disease, it provides for making up of the genomes. If any region of the genome is considered to be identical by descent, then the contiguous SNPs share the same genotype. When an affected individual is compared to the affected sibling, then all the identical regions are recorded. If the affected sibling and non- affected siblings do not possess the same disease phenotype [4], then it is clear that their DNA is different. Thus, it can be concluded that the IBD mapping results in removing any regions which are identical in both affected patients and unaffected sibling. This process is repeated for many families until it generates small, overlapping fragments, which contains the diseased gene.

e) Whole Exome Sequencing

Whole Exome Sequencing is an approach used in modern day sequencing technology and in the DNA sequence assembly tools for piecing together all the coding portions of the genome together [9]. The sequence is then compared to a reference genome and the differences inferred are noted. Except for pathogenic variants, all the polymorphisms, synonymous changes, and intronic changes are filtered and this technique is combined with other techniques to exclude pathogenic variants, which helps in identifying more diseases.

B. Computer Assisted Methods:

a) Fuzzy Clustering

It is mainly used for the formation of information granulation for solving the possible curse of dimensionality [2]. Principle Component Analysis (PCA) is the pre-processing tool which is used along with it for producing results. It gives better performance when it is used along with PCA than without using PCA in the datasets. On average, it was observed that about 50% of better improvement is having when PCA is used. It is observed from the table that the error rate becomes reduced when PCA is used.

Fuzzy Clustering	With use of PCA	Without the use of PCA
Optimal performance	11.03+0.452	0.088+0.033

Table 1: Fuzzy Clustering performance with or without PCA

b) Molecular Networks

Molecular networks are used to increase the prediction and also for finding the diseases from similar networks or the protein to protein

interaction [22]. The prediction of existence of diseases in the genes is improved by employing this molecular network [14, 22].

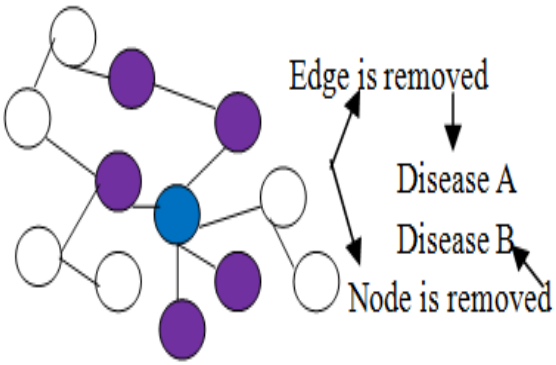

Fig.2: Molecular Network

The Blue circle indicates the Mutations in the node and the affected neighbour nodes by this mutation indicate in Purple circle. Disease A is the result of an edge removal, and disease B is from the node removal. These two diseases are not same, but it share similarity in their phenotypes [14]. By using this method, it will be easier to determine the existence of disease in the similar network structure.

a) Semi Non-Negative Matrix tri-Factorization

This method is used along with Co-clustering algorithm to classify the subtypes of the diseases such as cancer [3]. At present, this method is applied to create subtypes in cancer dataset but it will be give better results when applied to other diseases. The weights are assigned for each gene in the Gene Expression Profile and are accompanied with a network-assisted non-clustering algorithm for creating the subtypes of the disease. This method is applied to cancer database and it produces better results than other methods. But it remains tested to employ in other database other than the cancer diseases [3].

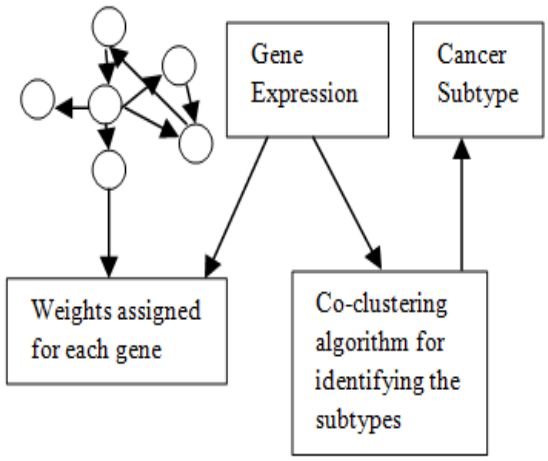

Fig.3: Working of Semi Non-Negative Matrix tri-Factorization with Co-Clustering Algorithm

c) Human Phenotype Ontology

This method is applied for capturing the phenotype similarities between the diseases that resides in the genes. This clinical information regarding OMIM has been generated in text mining. For example: 'Hypoplastic Philtrum' and 'Smooth Philtrum' are closely related to 'Hypoplastic nasal septum'. All these three diseases are comes under the category of NOSE of OMIM diseases [6]. Since all these terms are denoting the NOSE category of OMIM diseases but each disease are entirely different in nature [14].

The drawback that will be noted in this Ontology is that it mainly utilizes the text –mining approaches in which some indexing terms are not specifically designed to describe the human diseases [6]. But the structure and the terms of ontology based on the medical knowledge and it will be refined.

d) Co-relation Fractal Dimension

This method is applied in the research for selecting a suitable threshold value to find the group or the cluster. PSO algorithm is used along with this method to optimize the minimum threshold value [13].

Minimum Threshold Value	KDD dataset	Forest Dataset
	Purity	
0.20 0.21 0.22 0.23	} Yielded good results Maximum Value	

Table 2: Maximum Purity attained by setting Minimum Threshold Value

From the above table, it is observed that the minimum threshold value is set to 0.23 and at this point, maximum value of purity of both datasets is obtained [13]. From 0.2 to 0.23, better results produce and at the peak, the maximum value of purity attains at 0.23. This method along with PSO algorithm makes to set suitable threshold value.

But the drawback is that it only applied on the other fields which are not relevant to the medical field. It remains yet not tested in the medical field whether it is fit to all other fields or fit only to the forestry cover datasets [13].

III. DISCUSSION

Table III illustrates the comparison for computer assisted methods for the identifying the inherited

diseases with features such as Uses, Applications, Supported Algorithms, Datasets used, Database and Parameters used. All the methods are used only for Gene profiling and Co-relation Fractal Dimension is exemption to that.

Mostly all methods detect the existence of inherited diseases in the gene sequences and it produces the results more accurate and useful. Mostly all the mentioned methods are accompanied with supported algorithm to produce better results.

The Fuzzy Clustering applied over the datasets such as Treasury, Mortgage, and Weather other than the gene profiling datasets with the parameters like Accuracy, Prediction Capabilities, and Stability. In similar, the Co-relation Fractal Dimension is very commonly applied to statistical data such as forest cover type etc with parameters time, speed and horizon.

The limitations of Molecular Networks and Human Phenotype Ontology are it requires an optimized algorithm to detect the existence of inherited diseases in the genes. When it is activated with new optimized algorithm, its accuracy level becomes increases.

TABLE III: COMPARISON OF METHODS APPLIED IN HUMAN GENOME SEQUENCES

Method / Features	Fuzzy Clustering	Molecular Networks	Semi Non-Negative Matrix tri-Factorization	Human Phenotype Ontology	Co-relation Fractal Dimension
Uses	Overcome a Possible curse of dimensionality for the formation of information granulation	Find the Protein to Protein Interaction in the genes	Grouping similar samples of genes	Capturing the phenotype similarities between the diseases	Selecting the minimum Threshold value
Applications	Engineering, Medical Engineering, Social Science	Gene Expression Profiling	Gene Expression Profiling	Human Genome Sequences, Gene Mutations	Statistical Data
Supported Algorithm	HRBFNN	Shortest Path algorithm	Network – Assisted Co-clustering Algorithm	It doesn't activated with any other algorithm	PSO algorithm
Limitations	It gives better performance along with PCA otherwise it gives only 50% improvement	It has not activated with an Optimized algorithm	1.Network that have been used is not specified for particular type of Cancer 2.Network does not contains all the genes 3.Directions specified for the edges in the network are not clear	Most of the work used are text-mined concepts to map phenotype concepts	It is suited for Statistical data but it remains investigate whether it is fit to other data streams as well.
Datasets Used	Treasury, Mortgage, and Weather	Psoriasis	Breast Cancer, Brain Cancer	OMIM	KDD cup 1999, Forest cover type
Parameters Used	Accuracy, Prediction Capabilities, Stability	Accuracy	Survival Time, Tumor necrosis Percentage, Tumor Nuclei Percentage	Density, Rank	Speed, Time, Horizon
Database	Abalone data (ABA), MIS	BIND, HPRD, IntAct, DIP	TCGA	Medline, HPO	KDD

IV. CONCLUSIONS

The availability of Methods applied in Human Genome Sequences make researchers to quickly check new ideas in the forefront of the exciting and challenging field in genomic sequence. This paper discusses the vital role of computer assisted methods applied for detecting inherited diseases (i.e.) from the gene sequence. A comparative study of the existing methods reveals that each method is for a specific purpose accompanied with existing algorithms to produce better results. But, few methods need optimized algorithm to produce better result. Out of five methods compared, two methods not activated with any algorithm. If it is activated with any optimized algorithm, then it will give better results than the previous one. The Co-relation Fractal Dimension Method is known to be best suited for Statistical data but it remains investigate whether it is fit to other data streams. When it is applied in gene profiling concepts such that for detecting the existence of inheriting disease in the genes, then it will be tested that whether it is suited for all other datasets or not.

REFERENCES

1. N, Merikangas K. "The future of genetic studies of complex human diseases." Science 273: 1516-7, 1996
2. Wei Huang, Sung-kwun oh, and witold pedrycz, Design of hybrid radial basis function neural networks(HRBFNNS) realized with the aid of hybridization of fuzzy clustering method (FCM) and polynomial neural networks (PNNs),Neural Networks,60,166-181,2014
3. Yiyi Liu, Quanquan Gu, Jack P Hou, Jiawei Han and Jian Ma, A network assisted co-clustering algorithm to discover cancer subtypes based on gene expression, BMC Bioinformatics 2014
4. Veer, L.J.,Gene expression profiling predicts clinical outcome of breast cancer. Nature,415, 530-536,2002
5. Amb et al., "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease." Science, 313, 1929–1935. 86, (2006)
6. Peter N.Robinson, Sebastian Kohler, Sebastian Baver, Dominik Seelow,Denise Horn and Stefen Mundlos, The Human Phenotype Ontology: A Tool for Annotating and analyzing Human Hereditary Disease, The American Journal of Human Genetics, 83,610-615,2008
7. Yuhaizhao, Guoren Wang, Xiang Zhang, Jeffrey Xu Yu and Zhanghui Wang, Learning Phenotype structure Using sequence model, IEEE Transactions on Knowledge and Data Engineering, Vol.26, No. 3, 2014
8. Venter, J. C. et al.," The sequence of the human genome", Science 291, 1304–1351 (2001).
9. Lander E. S.," The new genomics: global views of biology." Science, 274, 536–539 (1996).
10. Timothy Sterne-Weiler, Jonathan Howard, Matthew Mort, David N-cooper and Jeremy R.Sanford,Loss of exon identity is a common mechanism of human inherited disease, Genome Research, 2014
11. Quan Zhong, Nicolas simonies,Qian-Ruli, Edgetic Perturbation models of human inherited disorders,Molecular Systems Biology,No.321,2009
12. Yong Lu,Roni Rosenfeld,Itamar Simon, Gerard J.Nau and Ziv Bar-Joseph, A probabilistic generative model for GO enrichment analysis, Nucleic Acids Research,Vol.36, No.17,2008
13. Igor Feldman, Andrey Rzhetsky and Dennis vitkup, Network properties of genes harboring inherited disease mutations, The National Academy of sciences of the USA, PNAS, No.11,Vol.105, 2008
14. I.Guyon, J. Weston, S. Barnhill, M.D and V. Vapni, Gene Selection for cancer classification using support vector machines. Machine Learning, 2000
15. Claustres M , Time for a unified system of mutation description and reporting: a review of locus specific mutation databases. Genome Res 12:680-688, 2002
16. Zoltan Dezso, Yuri Nilcosky,Craig webb, Identifying disease-specific genes based on their topological significance in protein networks, BMC Systems Biology. Pg.no: 1-14,2009
17. Usha Kuppusamy, Yanli Wang, Seshan Anathasubramanian "Predicting gene Ontology annotations of Orphan GWAS

genes using protein-protein interactions, Algorithms for Molecular Biology 2014,9:10

18. Xiujuan Wang, Natali Gulbahce and Haiyuan Yu,"Network based methods for human disease gene prediction", Briefings in functional genomics101.10.No.5, 280-293

19. Usha Kuppuswamy, Yanli Wang, Seshan Anantha subramanian, 'Predicting gene ontology annotations of orphan GWAS genes using protein-protein interactions', Algorithms for molecular Biology 2014,9:10

20. Anuradha Yarlagadda, J.V.R. Murthy, H.M.Krishna prasadd,'Particle Swarm Optimized optimal threshold value selection for clustering based on Correlation Fractal Dimension, Scientific Research, 2014,5,1615-1622

21. Yu-Ping Wang, 'Multiscale Genomic Imaging Informatics, Life Science – IEEE Signal Processing Magazine,169,2009

22. Madhavi.K.Ganapathiraju, Judith Klein-seetharaman, N.Balakrishnan and Raj Reddy, "Characterization of Protein Structure', IEEE signal Processing Magazine, 2004

23. Chun-Hsi Huang and Sanguthevar Rajasekaran, Parallal Pattern Identification in Biological Sequences on Clusters, IEEE Transactions on NanobioScience, Vol.2,No.1, 2003

24. Tieng K.Yap,Ophir Frieder and Robert L.Martino,'Parallel Computation in Biological Sequence.

25. Tieng K.Yap,Ophir Frieder and Robert L.Martino,'Parallel Computation in Biological Sequence Analysis', IEEE Transactons on Parallel and Distributed Systems, vol.9, No.3, 1998