Computational Identification of Biomarkers

Archana N. Mahajan, Maulika S. Patel

Abstract*: Early detection of diseases and personalized medicine are gaining huge attention as a result of the advancements in the fields of bioinformatics and computational biology. Computational Biology is an interdisciplinary area involving Biomolecular data understanding and analysis by using and development of various tools, algorithms and methods. Disease detection and prediction is mostly carried out through the symptoms reported by patient and clinical test performed on the basis of it. There is a need to detect disease before the symptoms progress. Biomarkers are biological markers which are implicitly available in biomolecular data. There are different types of biomarkers which can help in early disease detection and finding correlation of the disease with other changes at the cellular level. The biomolecular data exploration is the key to identify various biomarkers. This paper presents a summary of types of biomarkers and biomarker identification techniques.

***Keywords*: Biomarker, Computational biology, Decision Trees, Genomics, k-means clustering, Proteomics, Support vector machines, Transcriptomics**

I. INTRODUCTION

Health Informatics is information engineering applied to the field of health care, that is the management and use of patient healthcare information. Health informatics describes various methods to acquire, store, use of patient's healthcare data. It involves the interdisciplinary approach of the design, development, adoption and application of information technology-based innovations in healthcare services-delivery, management and planning [1][2].

Bioinformatics is the subset of health informatics which comprises of how different computational techniques can be used to analyze omics data sequences gathered through biological experiments and life sciences. The computational techniques are part of computational biology which mainly focuses on studying the biology through omics sequences such as genomic, proteomics, transcriptomics, and proteo-genomics. Omics data sequences have very large feature set with diverse categories. Bioinformatics and computational biology use several machine learning algorithms to deal with diverse categories of omics sequences, learn from these sequences and provide intelligent analysis for hypothesis formation and validation.

Personalized medicine is a systematic way to customize individual patient's treatment based on analyzing individual's omics sequences in a particular disease condition.

Personalized medicine not only helps in tailoring the medicine but also in pre-disease state identification and analysis, early detection of disease, active monitoring and management of treatment responses. Pre-disease state is the state in which there are significant molecular changes observed at a particular omics sequences at a particular time. The analysis of these changes help in early detection of disease.

The analysis of omics sequences is also able to provide detailed insight on finding association between genotype and phenotype as well as for identifying biomarkers for patient stratification[3]. In patient stratification the entire population of patient group is divided into subgroups based on feature similarity in a particular group and feature diversity among different subgroups.

II. OMICS DATA SEQUENCES

Advances in information technology lead to automation in various healthcare services. This generates diverse category and amount of medical data. The medical data in clinical health informatics comprises health signal monitoring, physiological data, radiology image data, ophthalmological data, omics data. The medical data in computational health informatics involves computational aspects to individual patient's care such as collection of data from medical equipment to electronic databases, statistical data analysis of disease diagnosis and progression data as well as patient signaling data.

Computational medical data involves hybrid combination of omics data sequences such as genomic, proteomic, radiomics, proteogenomic, proteoradiomic. Omics data sequences refer to the data in the organic and molecular fields such as genomics, proteomics, transcriptomics and radiomics. Genomics is the study of whole genomes of organisms; study of total DNA in cell/organism. Proteomics is the study of proteomes along with their structures and functions. A proteome is the entire set of proteins in a cell. Transcriptomics is the study of ribonucleic acid(RNA) sequences. Radiomics is the study of large number of features computationally extracted from radiological images in order to discover hidden features of certain region(s). Both clinical and computational data has an implicit type of data named biomarker data. A biomarker is a "defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention, including therapeutic interventions" [3].

A. Molecular Biomarkers

Molecular biomarkers are also known as omics biomarkers and have three distinct categories; Genomics, Transcriptomics, proteomics.

Revised Manuscript Received on May 20, 2020.

Archana N. Mahajan, PhD Scholar, Gujarat Technological University, Ahmedabad, Gujarat, India. E-mail: archanamagare@gmail.com

Dr. Maulika S. Patel, Professor & Head, Department of Computer Engineering, G H Patel College of Engg. & Tech., Vallabh Vidyanagar, Gujarat, India. E-mail: maulika.sandip@gmail.com

Genomic Biomarkers [4]: Genomic biomarker consists of one or more deoxyribonucleic acid (DNA) and its characteristics. Various DNA characteristics in genomic biomarkers are Single Nucleotide Polymorphism (SNP), short tandem repeats, haplotypes, DNA methylation. A single-nucleotide polymorphism (SNP) is a DNA sequence difference occurs when a single nucleotide adenine (A), thymine (T), cytosine (C), or guanine (G) in the genome changes between members of a species or paired chromosomes in an individual. A Short tandem repeat is a microsatellite, consisting of a unit of two to thirteen nucleotides repeated several to dozens of times in a row on the DNA strand. A haplotype is a set of DNA variations, or polymorphisms, that tend to be inherited together. DNA methylation is a process by which methyl groups are added to the DNA molecule. Methylation can change the activity of a DNA segment without changing the sequence.

Transcriptomics Biomarkers [4]: Transcriptomics biomarker consists of one or more RNA molecules within a cell and its characteristics. RNA characteristics are RNA sequences, RNA expression levels, RNA processing- e.g. splicing and editing. Transcriptomics biomarkers are discovered through Gene expression profiling at a particular biological state-Messenger RNA (mRNA), Gene expression profiling at a particular biological state- Noncoding RNA (ncRNA).

Proteomic Biomarkers [4]: Proteomic Biomarkers are discovered through protein expression, protein localization, protein- protein Interaction, post translational Modification.

B. Imaging Biomarkers

These are radiomics biomarkers which extract large amounts of quantitative and qualitative features from multimodality medical image. These biomarkers help in finding the correlations between the features and diagnosis/prognosis of disease.

C. Predictive Biomarkers

Predictive biomarkers[5] assess the most likely response to a particular treatment type for a particular disease. These biomarkers work as target for therapy.

D. Prognostic Biomarkers

Prognostic biomarkers[5] indicate progression of disease with or without treatment. It helps in monitoring different stages of the disease with appropriate level of specificity and sensitivity. Specificity highlights differentiating parameters in different disease stages. Sensitivity identifies notable changes in the test parameters of the disease.

E. Pharmacodynamics Biomarkers

Pharmacodynamics biomarkers[5] are molecular indicators of drug effect on the target in an organism. There are significant level changes observed at molecular parameters in response to the drug therapy changes in these biomarkers.

III. MACHINE LEARNING TECHNIQUES FOR HEALTH INFORMATICS

Biological experiments and life sciences generate voluminous data. Analysis of this data only with biological experiments is time consuming, costly and requires huge human resources. The bioinformatics deals with such data using various machine learning techniques.

The machine learning techniques used in bioinformatics are categorized into two broad categories as supervised and unsupervised learning. The supervised learning techniques are also known as classification techniques that map the input-output data based on previous labelled training data. The classification techniques involves two phases, learning phase and classification phase[6]**.** In learning phase, the classification model is built through the pre-classified examples. This phase is also known as training phase. In classification phase, the classification model built in the training phase is used to predict the class label for unknown instances. This phase is also known as the testing phase.

Classification methods used in literature are decision tree classification, random forest classification and support vector machine. Decision tree classifier [7] is used to understand several features and their correlation with labelled input features. Decision tree classifiers forms the tree like structure each internal node indicates the test on attribute, each branch represent the output of the test and each terminal node represents the class label. It is used for predicting splice site of protein, assigning protein function as well as predicting protein-protein interaction. The features used for classification by decision tree are gene expression correlation, shared cellular localization and genomic distance between interacting genes. Random forest classification is ensemble method which combines several base classifiers together which creates improved composite classification model. Random forest classifier [8] is used to increase the classification accuracy through votes of all base classifiers. It is used to classify high dimensional data by constructing many decision trees. In bioinformatics, random forest classifiers[9] are used for analysis of microarray gene expression data, mass spectrometry-based proteomics data, genome-wide association study, Protein-protein interaction prediction, and biological sequence analysis. Support vector machine finds the optimal hyper plane between various groups of data; increases inter group distance and reduces intra group distance. Support Vector Machine[10] is used in bioinformatics for prediction of protein secondary structure, muti-class protein folding recognition and prediction of human signal peptide cleavage site.

The unsupervised machine learning techniques are also known as clustering which infer the patterns in given data without reference to any labelled data. Thus in the unsupervised machine learning the number of classes to be learned are not known in advance. The unsupervised learning techniques group data based on the similarity found in the input instance’s parameters. Most common clustering techniques used in bioinformatics are k-means clustering, hierarchical clustering, model based clustering , and grid based clustering. K -means clustering[11] is used for analyzing the differentiating characteristics of Alzheimer disease patients and normal individuals. Hierarchical clustering[11] is used in bioinformatics when there are very large set of dimensions involved in the input data.

IV. LITERATURE SURVEY ON BIOMARKER DISCOVERY

Healthcare data is multi-scale medical data which requires use of machine learning techniques to explore such data.

The literature survey of various types of biomarker discovery is as shown in table I.

Table 1. Biomarker Discovery Techniques

Research Objective	**Key Findings-Biomarkers**	**Evaluation Parameters**
Identify genetic signatures in patients with IBDs and clarify the potential molecular mechanisms underlying IBD subtype[12]	G protein subunit gamma 11 (GNG11), G protein subunit beta 4 (GNB4), Angiotensinogen(AGT),Phosphoinositide-3-kinase regulatory subunit 3 (PIK3R3) and C-C motif chemokine receptor 7 (CCR7)	Disease Specific genes differentiating UC from CD based degree of interaction(>51) in PPI
Identify precise molecular mechanisms for dasatinib resistance[13]	COL1A2, ITGB4, ITGA2, LAMA3 and RAC1	Identified hug genes have a strong association with pancreatic adenocarcinoma ,prognosis based on Kaplan-Meier survival analysis
To identify critical genes and potential drugs in Colorectal cancer(CRC)[14]	AURKA,CNB1, CCNF,EXO1	Functional changes of DEGs are associated with DNA replication. Upregulated biomarkers in CRC samples than normal
To Check comparative applicability of deep learning methods in the field of quantitative proteomics for biomarker identification [15]	Increased Classification accuracy using DL	Recall,precision,accuracy,F_1 score
To discover New diagnostics Biomarkers of skin cancer Melanoma[16]	Peripheral skin information based biomarker	Precision
To evaluate performance of various machine learning classification techniques for classifying endometriosis and control samples using both 100 transcriptomics and methylomics data[17]	NOTCH3 biomarker(Protein coding genes) is identified by all the classifier. RASSF2 –Biomarker is a tumor suppressor gene.	TMM, Quantile normalization
To find association of alcohol and HCC through bioinformatics methods[18]	ACTG1 and TLR3 alcohol associated biomarkers	Pearson correlation analysis of co-expression was performed with a coefficient >0.85 and P<0.05 as the criteria for investigation. Kaplan-Meier survival analysis for prognostic analysis
To identify salivary biomarkers for alcohol dependence disorder[19]	Seven miRNAs as biomarkers expressed in AA AD patients, and five miRNA biomarkers as EA AD patients	AD prediction Accuracy based on Gini Index >72.2%
to identify objective blood biomarkers for pain for psychiatric patients[20]	MFAP3,blood gene expression biomarkers as predictive of pain particularly in females.	Convergent Functional Genomic Score>6
To evaluate whether patient's age is a biomarker for disease diagnosis[21]	Age is biomarker for disease diagnosis	PCC, Sensitivity, Specificity

To identify key biomarkers of Lung adenocarcinoma for understanding disease progression.[22]	AGER, SFTPC, FABP4, CYP4B1, WIF1 and GREM1, SPINK1, MMP1, COL11A1 and SPP1	P < 0 05
To identify new biomarkers such as cytokines, chemokines for predicting patient’s mortality with disease IE. To determine if cytokines, chemokines assessed at IE diagnosis can predict in-hospital death[23]	IL-15, CCL4 and CRP levels are prognostic biomarkers	Two sided p-value<=5 %, Gini Impurities
To investigate pre-transplant serum exosomes using proteomics to identify novel recipient biomarkers of PGD prior to heart transplant[24]	31 proteins with differential protein expression and predictive association with the development of PGD	ROC
To set up an automatic screening method combining both transcriptomic databases and support vector machine (SVM)-based pattern recognition to select biomarkers that can be used in predicting and preventing GDM[25]	6 genes are identified as biomarker of GDM.	receiver operating characteristic (ROC) curve analysis of SVM for sensitivity and specificity
To examine the utility of a mass spectrometry based blood serum protein biomarker test for detection of CRC[26]	a blood protein panel consisting of LRG1, EGFR, ITIH4, HPX, and superoxide dismutase 3 are significant biomarkers for detection of CRC	Receivers operating Curve

V. SUMMARY

The literature survey identifies different types of biomarkers with genomic, transcriptomic, proteomic sequences. Biomarker identification process used bioinformatics and machine learning approach. Various tools used during biomarker identification process through bioinformatics are GO, KEGGs, DAVID, STRING. The biomarker identification process based on literature involves two major techniques -biomarker identification through bioinformatics and through machine learning is as shown in Figure 1 and Figure 2.

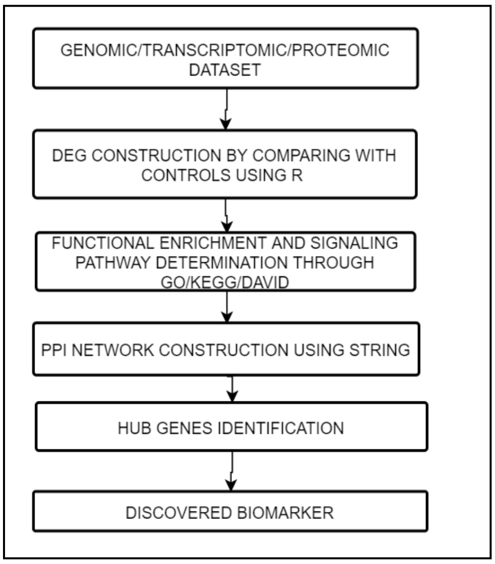
GENOMIC/TRANSCRIPTOMIC/PROTEOMIC DATASET
DEG CONSTRUCTION BY COMPARING WITH CONTROLS USING R
FUNCTIONAL ENRICHMENT AND SIGNALING PATHWAY DETERMINATION THROUGH GO/KEGG/DAVID
PPI NETWORK CONSTRUCTION USING STRING
HUB GENES IDENTIFICATION
DISCOVERED BIOMARKER

Fig.1 Biomarker Identification Brocess through Bioinformatics

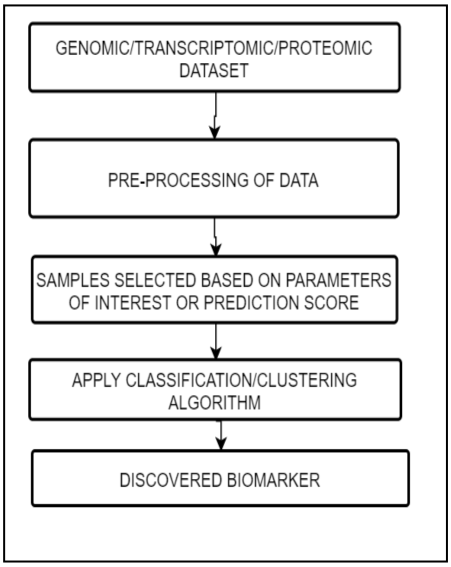
GENOMIC/TRANSCRIPTOMIC/PROTEOMIC DATASET
PRE-PROCESSING OF DATA
SAMPLES SELECTED BASED ON PARAMETERS OF INTEREST OR PREDICTION SCORE
APPLY CLASSIFICATION/CLUSTERING ALGORITHM
DISCOVERED BIOMARKER

Fig.2Biomarker Identification Process through ML

VI. DISCUSSION

Biomarkers are identified based on the different types of omics and image input data.

Various machine learning techniques and bioinformatics tools are used for biomarker identification to increase the accuracy. However, there is need to validate the identified biomarkers at subsequent level. There is need to integrate multi omics technologies for complete understanding of disease prognosis. Since multi omics data has heterogeneous sources, heterogeneous sample type and size, only bioinformatics tools are not adequate. Machine learning techniques has ability to handle multi omics data for biomarker identification and validation. Machine learning techniques in integration with bioinformatics tools can lead to biomarker validation.

Biomarker identification and validation can be carried out by integrating multi-omics technologies with hybrid analytical approach. Hybrid analytical approach is mixture of bioinformatics tools and machine learning techniques.

REFERENCES

1. Procter, R. (2009). Definition of Health Informatics.,from http://www.nlm.nih.gov/hsrinfo/informatics.html
2. Nadri H, Rahimi B, Timpka T, Sedghi S (August 2017). "The Top 100 Articles in the Medical Informatics: a Bibliometric Analysis", *Journal of Medical Systems. 41 (10): 150.* doi:10.1007/s10916-017-0794-4. PMID 28825158.
3. Strimbu K, Tavel JA. What are biomarkers?. Curr Opin HIV AIDS.2010;5(6):463–466. doi:10.1097/COH.0b013e32833ed177
4. Ich, “ICH Topic E15 Definitions for genomic biomarkers, pharmacogenomics, pharmacogenetics, genomic data and sample coding categories ,” *EMEA*, November, p. 1-8, 2007.
5. M. Gosho, K. Nagashima, and Y. Sato, “Study Designs and statistical analyses for biomarker research,” *Sensors (Switzerland)*, vol. 12, no. 7, pp. 8966–8986, 2012.
6. I. H. Witten, E. Frank, and M. a Hall, *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. 2011.
7. Carl Kingsford and Steven L Salzberg ,”What are decision trees?,Public Access,” *Bone*, vol. 23, no. 1, pp. 1–7, 2008.
8. E. Raczko and B. Zagajewski, “Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images,” *Eur. J. Remote Sens.*, vol. 50, no. 1, pp. 144–154, 2017.
9. Y. Qi, “Random forest for bioinformatics,” *Ensemble Mach. Learn. Methods Appl.*, pp. 307–323, 2012.
10. J.Y. Wang, “Applications of Support Vector Machines in bioinformatics,” *Bioinformatics*, pp. 1–56, 2002.
11. N. Toschi *et al.*, “Biomarker-guided clustering of Alzheimer’s disease clinical syndromes,” *Neurobiol. Aging*, vol. 83, pp. 42–53, 2019.
12. C. Cheng et al., “Identification of differentially expressed genes, associated functional terms pathways, and candidate diagnostic biomarkers in inflammatory bowel diseases by bioinformatics analysis,” *Exp. Ther. Med.*, pp. 278–288, 2019.
13. Jingsun Wei et al., ”Identification of biomarkers and their functions in dasatinib-resistant pancreatic cancer using bioinformatics analysis”,ONCOLOGY LETTERS 18: 197-206, April 2019
14. J. Chen, Z. Wang, X. Shen, X. Cui, and Y. Guo, “Identification of novel biomarkers and small molecule drugs in human colorectal cancer by microarray and bioinformatics analysis,” Mol. Genet. Genomic Med., vol. 7, no. 7, pp. 1–15, 2019.
15. H. Kim, Y. Kim, B. Han, J. Y. Jang, and Y. Kim, “Clinically Applicable Deep Learning Algorithm Using Quantitative Proteomic Data,” *J. Proteome Res.*, vol. 18, no. 8, pp. 3195–3202, 2019.
16. X. Li, J. Wu, E. Z. Chen, and H. Jiang, “From Deep Learning Towards Finding Skin Lesion Biomarkers,” *2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 2797–2800, 2019.
17. S. Akter *et al.*, “Machine Learning Classifiers for Endometriosis Using Transcriptomics and Methylomics Data,” *Front. Genet.*, vol. 10, 2019.
18. B. Gao, S. Li, Z. Tan, L. Ma, and J. I. A. Liu, “ACTG1 and TLR3 are biomarkers for alcohol-associated hepatocellular carcinoma,” *Oncol. Lett.*, vol. 17, no. 2, pp. 1714–1722, 2019.
19. A. J. Rosato *et al.*, “Salivary microRNAs identified by small RNA sequencing and machine learning as potential biomarkers of alcohol dependence,” *Epigenomics*, vol. 11, no. 7, pp. 739–749, 2019.
20. A. B. Niculescu *et al.*, “Towards precision medicine for pain: diagnostic biomarkers and repurposed drugs,” *Mol. Psychiatry*, vol. 24, no. 4, pp. 501–522, 2019.
21. X. Feng *et al.*, “Age is important for the early-stage detection of breast cancer on both transcriptomic and methylomic biomarkers,” *Front. Genet.*, vol. 10, no. MAR, pp. 1–13, 2019.
22. K. Ni and G. Sun, “The identification of key biomarkers in patients with lung adenocarcinoma based on bioinformatics,” *Math. Biosci. Eng.*, vol. 16, no. 6, pp. 7671–7687, 2019.
23. T. Ris *et al.*, “Inflammatory biomarkers in infective endocarditis: machine learning to predict mortality,” *Clin. Exp. Immunol.*, vol. 196, no. 3, pp. 374–382, 2019.
24. N. Giangreco *et al.*, “Exosome Proteomics and Machine Learning Identify Novel Biomarkers of Primary Graft Dysfunction,” *J. Hear. Lung Transplant.*, vol. 38, no. 4, p. S137, 2019.
25. Y. Wang, Z. Wang, and H. Zhang, “Identification of diagnostic biomarker in patients with gestational diabetes mellitus based on transcriptome-wide gene expression and pattern recognition,” *J. Cell. Biochem.*, vol. 120, no. 2, pp. 1503–1510, 2019.
26. M. M. Ivancic *et al.*, “Noninvasive Detection of Colorectal Carcinomas Using Serum Protein Biomarkers,” *J. Surg. Res.*, vol. 246, no. 205, pp. 160–169, 2019.