IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 11, Issue 10, October 2023

DOI: 10.17148/IJIREEICE.2023.111001

A Literature review on Biomarker discovery

Manila M V[1]

Lecturer, Computer Hardware Engg, Government Polytechnic College, Palakkad, India

Abstract: This paper presents a literature review on the Biomarker discovery. The advent of omics technologies like genomics, proteomics, transcriptomics, and metabolomics helped the researchers to discover novel biomarkers that can be used to diagnose, predict, and monitor the progress of disease. A number of computational approaches like machine learning and deep learning have been developed to identify biomarkers by using omics data. The integration of different omics known as multi omics has also been used by researchers for biomarker discovery. This paper presents a comprehensive survey of the recent progress in the identification of biomarkers.

Keywords: Biomarkers, Multi omics data, Machine learning, Deep learning.

I. INTRODUCTION

According to the NIH (U. S. National Institutes of Health), A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention[1]. Biomarkers can be found in tissue, blood, or other body fluids and can provide information about a variety of bodily functions and disease states. With the invent of precision medicine, molecular biomarkers have been extensively used for accurate diagnosis or prognosis. Biomarkers can be used clinically to screen for, diagnose, or monitor the activity of diseases and to guide molecularly targeted therapy or assess therapeutic response. Biomarkers can be classified based on different parameters. Biomarkers can be classified based on their characteristics, such as molecular biomarkers (DNA, RNA, genes, proteins, metabolites, etc.), image biomarkers (X-ray, computed tomography (CT), magnetic resonance images, positron emission tomographies, etc.).

There are five types of biomarkers, including diagnostic biomarkers (determining disease presence or subtypes), prognostic biomarkers (identifying likelihood of a clinical event, disease recurrence or progression), predictive biomarkers (identifying individuals who are more likely than similar individuals without the biomarker), monitoring biomarkers (assessing disease status, medical condition), and safety biomarkers (indicating the likelihood, presence of toxicity). With the advent of high throughput technologies a high amount of omics data like genome, proteome transcriptome and metabolome have been created. The whole sequence of DNA in an organism, including all of its chromosomes, is referred to as a genome. The entire universe of proteins in the cell is called proteome. A transcriptome is a collection of mRNA, miRNA, and lncRNA molecules in which their sequence produced in a particular cell is called “transcriptome. The metabolome contains a complete collection of smallmolecule groups called metabolites, including carbohydrates, amino acids, sugars, and fatty acids. The omics data have been widely used to identify biomarkers for diagnosis and prognosis of diseases. Sometimes it is impossible to identify biomarkers using single type of omics data[2]. So integrated omics or multi-omics is required for the discovery of biomarkers. The integrated omics is complex and can be handled by computational methods like machine learning and deep learning approaches. Machine learning and deep learning can be applied to single omics and multi-omics data for biomarker identification. In machine learning, biomarker identification can be treated as feature selection or gene prioritization whereas diagnosis can be regarded as classification problem while prognosis can be treated as a regression or classification problem. Numerous machine-learning approaches have been proposed by researchers for the identification of biomarkers for diseases. In this survey, we present a comprehensive review of the identification of biomarkers with machine learning approaches.

II. LITERATURE SURVEY

In the past few years, a huge amount of omics data have been deposited into public databases with the advances in high-throughput technologies, such as the Cancer Genome Atlas (TCGA), the Human Protein Alas (HPA), the Catalogue Of Somatic Mutations In Cancer (COSMIC). In recent years genomics and transcriptomics data are the most common omics data available in public databases due to the decrease in sequencing costs. Circular RNAs are identified as diagnostic markers in Hepatocellular carcinoma patients [3]. Long noncoding RNA (lncRNA) are recognized as diagnostic and prognostic markers in gastric cancer [4]. These omics resources are used to identify signatures for disease diagnosis and prediction. Sometimes, it is impossible to identify biomarkers using a single type of omics data [2]. Integrated omics data are used to identify diagnostic, prognostic, and predictive biomarkers is suggested in [5].

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 11, Issue 10, October 2023

DOI: 10.17148/IJIREEICE.2023.111001

The identification of biomarkers can be regarded as a feature selection problem. The omics data have high dimensionality and noise which make it difficult to extract data from it. In machine learning, there are two methods used for dimensionality reduction, a) feature extraction and b) feature selection. Feature Extraction reduces the feature space of high dimensional multi-omics data to low dimensional feature space[6]. This low-dimensional feature space can be used for biomarker identification. The dimensionality reduction techniques can be classified into linear and non-linear approaches. There are different techniques for the extraction of features for integrated linear multi-omics data, including Principal Component Analysis (PCA) [7], Canonical Correlation Analysis (CCA)[8], and Nonnegative Matrix Factorization (NMF) [9]. PCA is used for dimensionality reduction for both single omics and multi-omics data, which is good for normal distributed data, and may fail to work if the data distribution is strongly skewed. CCA is used for detecting the correlation between two or more datasets by transforming distinct datasets into different new spaces so that these data can be maximally correlated. NMF is used for integrative analysis of multi-omics data with low dimensional representation of the original matrix. The nonlinear approaches for nonlinear integrative analysis of multi-omics data include kernel principle component analysis (KPCA) [10], kernel canonical correlation analysis (KCCA) [11], Locally Linear Embedding (LLE)[12], t-distributed Stochastic Neighbor Embedding (t-SNE) [13], and auto-encoders [14]. Feature extraction can return only a subset of features, but some relevant features need to be selected for biomarker identification, which is done using feature selection. Feature selection is a method of electing valuable and informative features by removing duplicate and noisy features. Feature selection techniques are of three types, including filter, wrapper, and embedded methods. Different filter method techniques include Pearson Correlation Coefficient (PCC), chisquare, t-test, and Analysis of Variance, which work by finding the correlation between the features and target variable. In biomarker identification, chi-square and t-test are used by various researchers to rank the differentially expressed genes and select the top-ranked genes. Different wrapper methods are Recursive Feature Elimination (RFE), Sequential Feature Elimination (SFE), and Genetic Algorithms. The embedded method combines the function of both the filter and wrapper methods. Once the biomarkers are identified using feature selection and extraction, they are passed to machine and deep learning algorithms for the classification of biomarkers. The complete workflow of biomarker identification using Machine learning and Deep learning is shown in the below figure.

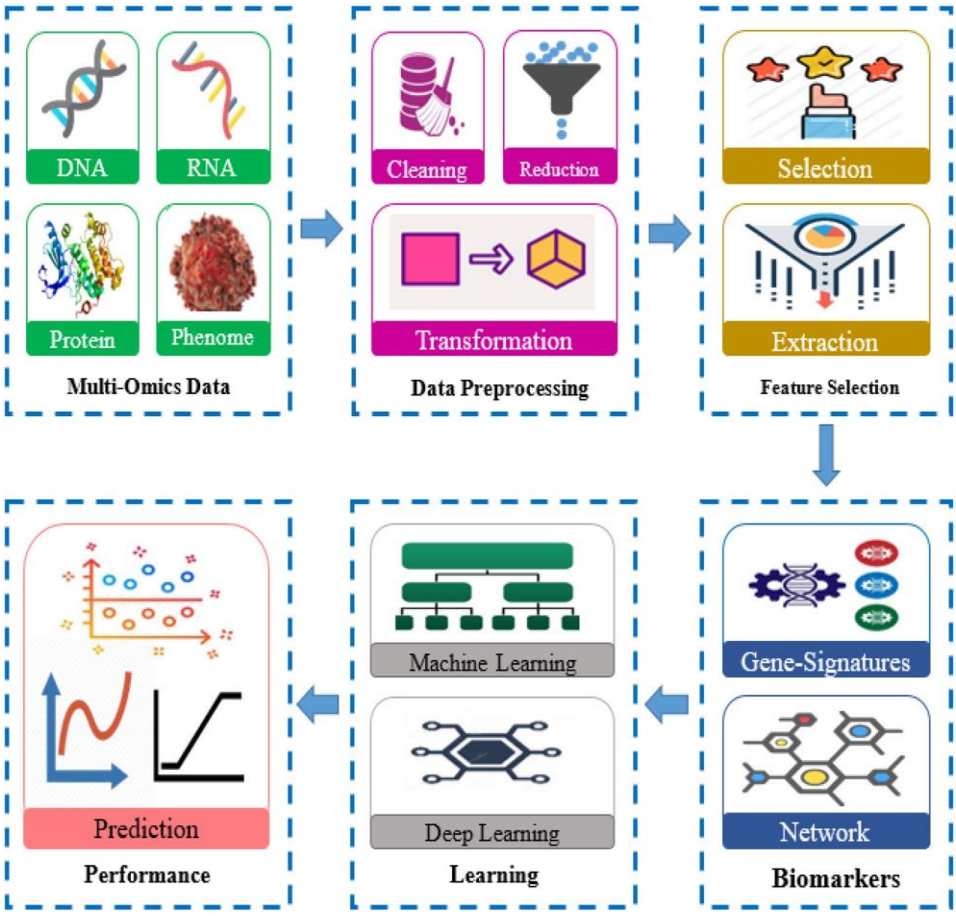

Fig. 1. Workflow of Biomarker identification [15]

DOI: 10.17148/IJIREEICE.2023.111001

III. CONCLUSION

The collection of different forms of omics data in the postgenomics period allows for the screening of specific markers for accurate diagnosis and prognosis, which is essential in personalized medicine. Unfortunately, identifying biomarkers from a large volume of omics data, particularly when there are complex interactions between molecules, is a difficult task. From the past research work, it is found that single type of data is not enough for identification of genes in patients. Therefore, multi-omics data is required for accurate discovery of markers and to guide treatment therapies based on the identified markers.

REFERENCES

[1]. Biomarkers Definitions Working Group., "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework," Clinical pharmacology and therapeutics, vol. 69, p. 89—95, March 2001.
[2]. H. Cao and E. Schwarz, "Opportunities and challenges of machine learning approaches for biomarker signature identification in psychiatry," Personalized Psychiatry, pp. 117–126, 2020.
[3]. K. Zhu, H. Zhan, Y. Peng, L. Yang, Q. Gao, H. Jia, Z. Dai, Z. Tang, J. Fan, and J. Zhou, "Plasma hsa circ 0027089 is a diagnostic biomarker for hepatitis b virus-related hepatocellular carcinoma," Carcinogenesis, vol. 41, no. 3, pp. 296–302, 2020.
[4]. S. Fattahi, M. Kosari-Monfared, M. Golpour, Z. Emami, M. Ghasemiyan, M. Nouri, and H. Akhavan-Niaki,"Lncrnas as potential diagnostic and prognostic biomarkers in gastric cancer: A novel approach to personalized medicine," Journal of Cellular Physiology, vol. 235, pp. 3189 – 3206, 2020.
[5]. K. Shi, W. Lin, and X.-M. Zhao, "Identifying molecular biomarkers for diseases with machine learning based on integrative omics," IEEE/ACM transactions on computational biology and bioinformatics, vol. 18, no. 6, pp. 2514–2525, 2020.
[6]. A. Dhillon and A. Singh, "ebrecap: Extreme learning based model for breast cancer survival prediction," IET Systems Biology, vol. 14, 05 2020.
[7]. M. Ringner, "What is principal component analysis?," ´ Nature Biotechnology, vol. 26, pp. 303–4, 04 2008.
[8]. D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," Neural computation, vol. 16, pp. 2639–64, 01 2005.
[9]. S. H. S. Lee, Daniel D, "Learning the parts of objects by non-negative matrix factorization," Nature, pp. 788–791, 10 1999.
[10]. B. Scholkopf, A. Smola, and K.-R. M ¨ uller, "Nonlinear component ¨ analysis as a kernel eigenvalue problem," Neural Computation, vol. 10, pp. 1299–1319, 07 1998.
[11]. P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," International journal of neural systems, vol. 10, no. 05, pp. 365–377, 2000.
[12]. "Supervised locally linear embedding with probability-based distance for classification," Computers Mathematics with Applications, vol. 57, no. 6, pp. 919–926, 2009.
[13]. L. van der Maaten and G. Hinton, "Viualizing data using t-sne," Journal of Machine Learning Research, vol. 9, pp. 2579–2605, 11 2008.
[14]. "Auto-encoder based dimensionality reduction," Neurocomputing, vol. 184, pp. 232–242, 2016.
[15]. A. Dhillon and A. Singh, "A systematic review on biomarker identification for cancer diagnosis and prognosis in multi-omics: From computational needs to machine learning and deep learning," Archives of Computational Methods in Engineering, 09 2022.