# SCRUTINIZING THE DISEASE BASED ON OMICS

DEPARTMENT OF INFORMATION TECHNOLOGY

PROJECT GUIDE

MR.N.BALA SUNDARA GANAPATHY
ASST.PROFESSOR GRADE I,
DEPARTMENT OF IT,
PANIMALAR ENGINEERING COLLEGE,
CHENNAI – 123.

PRESENTED BY

M.SATHISH
C.SARAVANAN
G.SETHU PANDIAN
P.INDIRAJITH

isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology 2

Volume: 6, Issue: 3,Special Issue: 1 ,Mar,2021 ,ISSN_NO: 2321-3337

# Scrutinizing the disease based on omics

**Abstract— Molecular biomarkers are certain molecules or set of molecules that can be of help for diagnosis or prognosis of diseases or disorders. In the past decades, thanks to the advances in high-throughput technologies, a huge amount of molecular 'omics' data, e.g. transcriptomics and proteomics, have been accumulated. The availability of these omics data makes it possible to screen biomarkers for diseases or disorders. Accordingly, a number of computational approaches have been developed to identify biomarkers by exploring the omics data. In this review, we present a comprehensive survey on the recent progress of identification of molecular biomarkers with machine learning approaches. Specifically, we categorize the machine learning approaches into supervised, un-supervised and recommendation approaches, where the biomarkers including single genes, gene sets and small gene networks. In addition, we further discuss potential problems underlying bio-medical data that may pose challenges for machine learning, and provide possible directions for future biomarker identification.**

**Index Terms—Molecular biomarker, machine learning, precision medicine, disease diagnosis, gene prioritization**

## 1 INTRODUCTION

Over the past decades, major efforts have been made for the treatment and prevention of complex diseases. Especially, with the launch of precision medicine, molecular biomarkers have been extensively used for accurate diagnosis or prognosis [1]. For example, mutations in the gene *SAMHD1* are highly associated with the development of malignancies, including cutaneous T cell lymphoma, chronic lymphatic leukemia [2] and colon cancer [3]. Recently, the gene has been used as biomarker and therapeutic target for acute myeloid leukemia [4]. Except for protein-coding gene, non-coding genes, e.g. circular RNAs (circRNAs) are emerging as biomarkers for diagnosis of diseases. For instance, F-circEA, a fusion circRNA, is recently reported to be a novel "liquid biopsy" biomarker of non-small cell lung cancer (NSCLC) by Tan *et al.* [5]. These biomarkers are valuable for diagnosis or prognosis of diseases.

In literature, biomarkers are defined as objectively measurable and evaluable indicators for normal biological process, pathogenic processes or responses to a therapeutic intervention [6]. For disease, the biomarkers are those that can distinguish disease state from normal state, or separate disease stages. There are many types of biomarkers for diseases, such as molecular biomarkers (DNA, RNA, genes, proteins, metabolites, etc.), image biomarkers (magnetic resonance images, positron emission tomographies, etc.), and so on. According to definitions of biomarkers from the BEST Resource [7], 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). In this survey, we will focus on diagnostic, prognostic and predictive biomarkers for diseases, and only molecular biomarkers are considered here.

Recently, thanks to the advances in high-throughput technologies, a huge amount of molecular 'omics' data, e.g. transcriptomics and proteomics, have been accumulated. Despite that the availability of the omics data makes it possible to screen biomarkers for diseases or disorders, it is a big challenge to identify biomarkers that can accurately diagnose or predict diseases considering tens of thousands of genes and millions of mutations in the omics data. As shown in Fig. 1, in machine learning, diagnosis can be regarded as classification problem while prognosis can be treated as regression or classification problem, where biomarker identification can be treated as feature selection or prioritization. Accordingly, a number of machine learning approaches have been proposed for identification of molecular biomarkers for diseases. In this

This work was partly supported by National Natural Science Foundation of China (61932008, 61772368) National Key R&D Program of China (2018YFC0910500), Natural Science Foundation of Shanghai (17ZR1445600), Shanghai Municipal Science and Technology Major Project (2018SHZDZX01), ZJLab and the Fundamental Research Funds for young teacher of Guangxi (2017KY0264).

Kai Shi is a postdoctor at the School of Mathematical Sciences, Fudan University, Shanghai 200433, China, and also with the College of Science, Guilin University of Technology, Guilin, Guangxi, 541004, China, (e-mail: mail_shikai @ foxmail.com).

Wei Lin is with the Institute of Science and Technology for Brain-Inspired Intelligence (ISTBI), Fudan University, Shanghai 200433, China, (e-mail: wlin@fudan.edu.cn).

Xing-Ming Zhao is with the Institute of Science and Technology for Brain-Inspired Intelligence (ISTBI), Fudan University, Shanghai 200433, China, and also with Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Ministry of Education, China, (e-mail: xmzhao@fudan.edu.cn). (Corresponding author: Xing-Ming Zhao.)

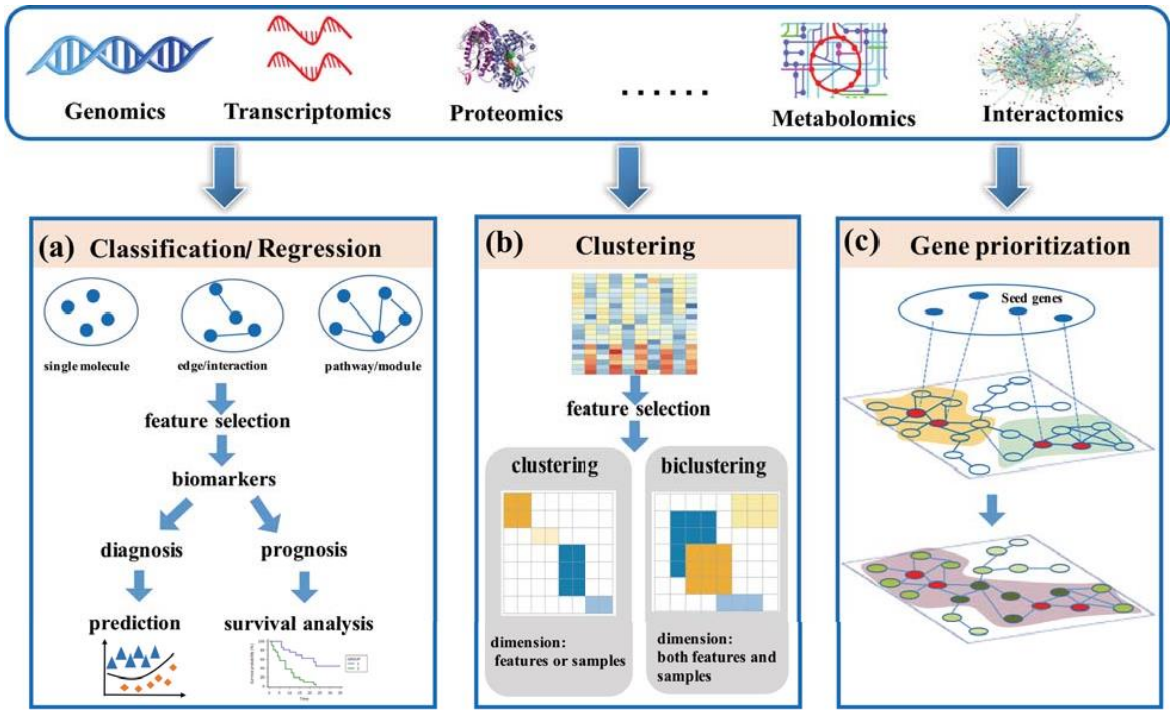



Fig. 1. The identification of molecular biomarkers with machine learning. (a) In supervised learning, diagnosis is treated as classification and prognosis as regression problems, where biomarker identification is regarded as feature selection; (b) In un-supervised learning, subtype stratification is regarded as clustering problems, where sets of genes are used as biomarkers; (c) In recommendation systems, the identification of predictive biomarkers for diseases is regarded as gene prioritization problem, where the genes can be indicative of the occurrence of diseases.

survey, we present a comprehensive overview on the recent progress of identification of molecular biomarkers with machine learning approaches. Specifically, we categorize those machine learning approaches into supervised, un-supervised and recommendation approaches considering the problems to be faced. The molecular biomarkers considered here include gene biomarkers, interaction biomarkers and network biomarkers, that can be used for monogenic or polygenic diseases. In addition, we further discuss potential problems underlying bio-medical data that may pose challenges for machine learning, and provide possible directions for future biomarker identification.

The paper is organized as follows. Section 2 gives an overview of popular omics data resources. Section 3 introduces feature extraction and feature selection. Section 4 presents supervised learning approaches for identifying diagnostic and prognostic biomarkers. Section 5 presents disease biomarker identification with unsupervised learning approaches. Section 6 introduces predictive biomarker identification approaches with molecular networks, where the genes can be used for predicting the occurrence of diseases. Section 7 discusses different biology data integration and new approaches application on biology medicine. Finally, future perspectives of biomarker identification with machine learning are presented.

## 2 Popular Omics Resources

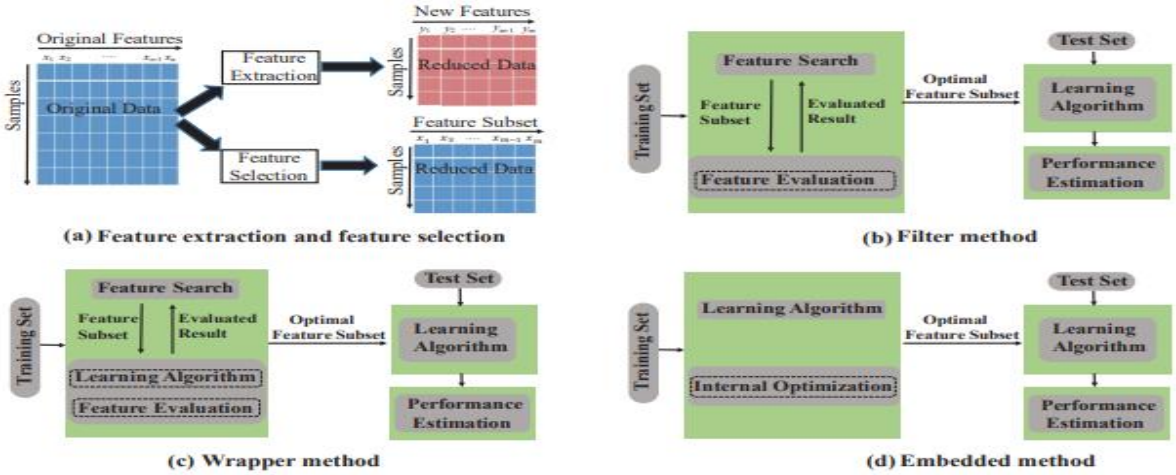
In recent years, a huge amount of molecular omics data have been deposited into public databases with the advances in high-throughput technologies, such as the Cancer Genome Atlas

(TCGA) [8], the Human Protein Alas (HPA) [9], the Catalogue Of Somatic Mutations In Cancer (COSMIC) [10]. Table. I shows the most popular resources of various kinds of molecular omics data that are widely used for identification of biomarkers[11, 12]. Among the omics data, genomics and transcriptomics data are the most common data available in the public databases due to decreasing of sequencing cost in recent years. The genomics data can provide the mutations or structural variations occurring in diseases, and can therefore identify predictive biomarkers for diseases. The transcriptomics data quantifies gene expression, and can help pinpoint genes that are aberrantly expressed in diseases. The proteomics data aims to quantify protein abundance, modification and interactions, and the metabolomics data seeks to identify and quantify the metabolites. The interactomics presents the landscape of molecular interactions in biological systems. The omics data have been widely used to identify biomarkers for diagnosis andprognosis of diseases.

## 3 Feature Extraction And Selection

In general, there are thousands or even tens of thousands of molecules considered in the omics data, where the high dimensionality and noise inherited in the data make it a big challenge to extract signals from the data. In machine learning, there are two common types of methods used for dimensionality reduction, i.e. feature extraction and feature selection, as shown in Fig. 2a.

*Feature extraction:* the feature extraction techniques have shown excellent performance for dimension reduction by transforming the original data into a lower dimensional space. In other words, new features generated by feature extraction are of the functions of the original features. Then, the new features are used as input of machine learning algorithm. The popular dimension reduction techniques can be grouped into linear and non-linear approaches. For example, principle component analysis (PCA) [13], a popular linear dimension approach, is widely used for dimensionality reduction for both single omics and multi-omics data, where the new features (principal components) are generally used as input for classification or clustering [14, 15]. Usually, PCA performs very well for the approximately normal distributed data, and may fail to work if the data distribution is strongly skewed. Canonical correlation analysis(CCA) [16] is especially useful 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. For example, the variant of PCA is used for disease subtyping with joint dimension reduction of multi-omics data [17], and the sparse version of CCA is used to screen markers of cardiovascular diseases for integrative analysis of transcriptomic and metabolomic data [18]. Nonnegative matrix factorization (NMF) [19] is another widely used approach for integrative analysis of multi-omics data with low dimensional representation of original data matrix. For




(a) Feature extraction and feature selection

(b) Filter method

(c) Wrapper method

(d) Embedded method

instance, NMF has been successfully used to integrate DNA methylation, gene expression and miRNA expression data in the identification of subtypes of ovarian cancer [20]. The linear approaches generally work well but may fail to handle the data with nonlinear structures, e.g. the interactome data.

The nonlinear approaches have been developed for non-linear integrative analysis of multi-omics data. The popular approaches include kernel principle component analysis (KPCA) [21], kernel canonical correlation analysis (KCCA) [22], and some manifold learning methods including isometric feature mapping (ISOMAP) [23], locally linear embedding (LLE) [24], Laplacian eigenmaps (LE) [25] and *t*-distributed stochastic neighborhood embedding (tSNE) [26]. These approaches generally map the original data into low-dimensional representations with non-linear transformations. For example, Mariette *et al.* [27] explored heterogeneous data integration, which used KPCA and kernel Self-Organizing Maps to cluster breast cancer patients with the integrative analysis of mRNA expression, miRNA expression and DNA methylation data. Recently, with the popularity of deep learning techniques, different variants of autoencoder [28] have been used to integrate different omics data. For example, Xu *et al*.

[29] integrated gene expression, miRNA expression and DNA methylation data to classify cancer subtype, where each omics data had a learned high-level representation with stacked autoencoder and all learned representations were further integrated and used for classification.

By transforming the original data into low dimensional representations, feature extraction can help extract useful signals from the original data and reduce the computation burden. However, the limitation of feature extraction is also obvious. For example, which feature extraction approach should be used for the data on hand and how many dimensions should choose from the new feature space. In addition, in somecases, it is difficult to explain why and how the new features contribute to the good performance of the downstream learningalgorithms.

*Feature selection:* it is an effective and efficient technique to reduce high-dimensional data, where a subset of informative features will be selected by removing redundant and noisy features. The features selected in this way can help explore how each feature performs and interpret why some features can improve the performance. The feature selection approaches are generally grouped into three categories, i.e. filter, wrapper and embedded approaches [30]. The filter methods (Fig. 2b) select features based on the association between features and class labels, which reflect the intrinsic characteristics of data. Usually, a filter method performs two steps: (1) ranks the features based on evaluation criteria and (2) filters the features with low ranking. For example, the *t*-test has been widely used to rank the differentially expressed genes or other molecules when discriminating cancers from controls [31-34]. The filter approaches are simple and easy to interpret, and are more suitable for high-dimensional omics data. However, the filter approaches assume the features to be independent and ignore the dependencies among features, which may be not reasonable for omics data considering the complex functional relationships between molecules. Furthermore, the filter approaches select features independent of classifiers, which means the features selected may be not the optimal ones for the classifier.

The wrapper methods (Fig. 2c) select features based on their performance during classification with certain classifier. Given a learning algorithm, a typical wrapper method consists of two steps. The first step is to search a subset of features based on a search strategy, and the second step is to evaluate the selected features based on classification error rate or performance.

accuracy. Then, it repeats the first step and the second step until the most discriminative feature subset appears. The wrapper methods are broadly applied to identify disease biomarkers with different search strategy. For example, the sequential forward selection has been used for screening risk genes [35] while sequential backward selection is used for screening relevant CpGs [34]. The wrapper methods take into account the interaction between the classifier and selected features, and thus generally achieve better performance than filter methods. However, this kind of method is more prone to risk of overfitting especially for datasets with small number of samples.In addition, higher computation cost may be required for the wrapper methods.

Unlike filter and wrapper methods, the embedded methods (Fig. 2d) integrate feature selection and classifier training together, and the feature selection is embedded in the learning process. Finally, the fitted model and selected features are obtained simultaneously. For example, least absolute shrinkage and selection operator (LASSO) is a popular embedded feature selection method for diagnosis markers and prognosis markers, where it is not only applied to homogeneous feature sets but also to heterogeneous concatenated feature sets form multiomics data [36, 37]. Compared with wrapper methods, the embedded methods have a low risk of overfitting, but may lead to higher computation burden for high-dimensional data.Compared with feature extraction, the features obtained with feature selection are easy to interpret without changing the original features. However, the signal hidden in the original feature space may be difficult to dig out. Therefore, feature extraction and selection can be used together in some cases, where new features are first extracted from the original data andthen a subset of new features will be selected. For example, Zhang *et al*. identified microbial biomarkers of obesity and metabolic syndrome, where the principal components (PCs) from gut microbial species data were first extracted and the bestcombination of PCs further selected by genetic algorithm was used as the input of the prediction model [38].

**EXISTING SYSTEM**

Over the past decades, major efforts have been made for the treatment and prevention of complex diseases. Especially, with the launch of precision medicine, molecular biomarkers have been extensively used for accurate diagnosis or prognosis. For example, mutations in the gene SAMHD1 are highly associated with the development of malignancies, including cutaneous T cell lymphoma, chronic lymphatic leukemic and colon cancer.Recently, the gene has been used as biomarker and therapeutic target for acute myeloid leukaemia. Except for protein-coding gene, non-coding genes, e.g. circular RNAs (circRNAs) are emerging as biomarkers for diagnosis of diseases. For instance, F-circEA, a fusion circRNA, is recently reported to be a novel “liquid biopsy” biomarker of non-small cell lung cancer (NSCLC) by Tan et al.. These biomarkers are valuable for diagnosis or prognosis of diseases.

**ABOUT THE PROJECT**

In this project, the main process is finding the patient disease by analysing the omics of each patient. Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms. Each structure of the disease is monitored automatically by analysing the structure and records of every patient. Biomarkers will do the treatment and identify the omics. In this higher will be updating the details previously (omics, medicine). The user will be uploading their result to the higher; testing is done for every patient – blood testing and temperature testing. Each testing has high, low, medium range which examines the patient readings. The biomarkers will be recording all results and by using omics it will automatically predict the patient condition.

**DISADVANTAGES**

It is difficult to learn and predict the disease of each patient.Structure of each disease cannot be verified easilyThere is no detailed format for each patient.The automation process is done in this biomarker.

**PROPOSED SYSTEM**

In machine learning, the diagnosis or subtyping of diseases is actually a classification problem, where the diagnosis is a binary classification problem while the subtyping is a multi-classification problem. In both diagnosis and subtyping, the identification of biomarkers can be regarded as a feature selection problem, where the biomarkers are those most informative features that can discriminate diseases from controls or classify disease samples into distinct subtypes. On the other hand, the prognosis of diseases is actually a regression problem in machine learning, where biomarkers are those molecules that are most associated with disease outcomes.In the following parts, the supervised learning approaches that have been proposed for identifying disease biomarkers are introduced.In particular, by taking into account the context of molecules of interest, the molecular biomarkers are grouped into single molecule biomarker, interaction biomarker, pathway biomarker, and network biomarker.

**ADVANTAGES**

The data of each patient is easily categorized by their details.The omics of each patient are monitored and updated automatically to the higher official.The bio markers will predict what kind of disease had affected the patient.The higher will update the proper medicine for the patient.Easy to calculate, and learn the every patient details automatically.It takes less time to get result from the higher official.

,

.

## BOTTOM LINES AND FUTURE ENCHANCEMENT

The main purpose of the project is used to store a large amount of data securely stored, the data are stored in the encrypted type. The data which are handled by the admin of the particular brand company used to upload a large amount of data regarding the formulation of the chemical products. In the future, can be built with many filters. In the future several brand products chemical formulation, the product ingredients were encrypted and the owner of the data handling only can access the application as the user. More details can be stored in this application

## HARDWARE AND SOFTWARE REQUIREMENTS

*Hardware Requirements*

- RAM 4GB
- Dual-Core 2.8 GHz Processor and Above
- HDD 80 GB Hard Disk Space and Above

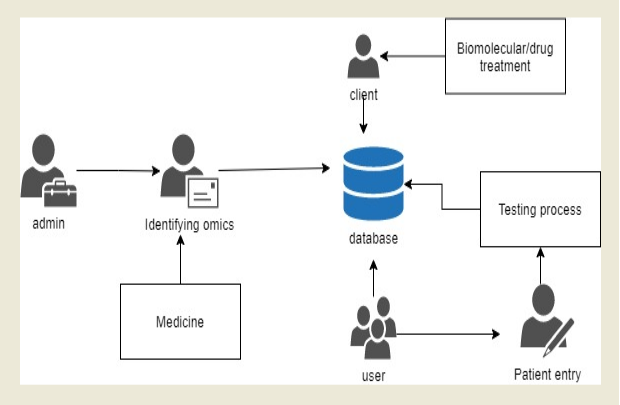
*Software Requirements*

- WINDOWS OS (7 /XP and Above)
- Visual Studio .Net 2015
- Visual Studio .Net Framework 4.5
- SQL Server 2014

## SCOPE OF THE PROJECT

The branches of science known informally as omics are various disciplines in biology whose names end in the suffix *-omics*,such as genomics, proteomics, metabolomics, and glycemic. Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms. Databases (both relational and otherwise) are a pretty important part of the computing experience. Modern systems make vast use of databases and their accompanying query technology to power just about every software application we depend on. Because these databases often contain sensitive information, there has been a strong push to *secure* that data. A key goal is to encrypt the contents of the database so that a malicious database operator (hacker) can't get access to it if they compromise a single machine. Many "omes" beyond the original "genome" have become useful and have been widely adopted by research scientists. "Proteomics" has become well-established as a term for studying proteins at a large scale. "Omes" can provide an easy shorthand to encapsulate a field; for example, an interact omics study is clearly recognizable as relating to large-scale analyses of gene-gene, protein-protein, or protein-ligand interactions.

## ARCHITECTURE DIAGRAM




## LIST OF MODULES

- *Testing process*
- *Biomarkers identification*
- *Omics identification*
- *Biomolecules characterization*

**MODULE EXPLANATION**

***MODULE NAME*** **:** Testing process

**BRIEF DESCRIPTION**

The patient will have general meeting they will be collecting the information of the user (patient). Oral information of each patient is collected by the higher. In this each patient has the testing process, to check their blood and temperature characterization, the user will be updating their blood test and temperature test. The data of testing level will be compared with the omics biomolecule.

***MODULE NAME*** **:** Biomarkers Identification

**BRIEF DESCRIPTION**

The biomarker process is the drug treatment, it will identify the each biomolecules of the patient's testing level, and they will predict the data comparing with the omics structure. The feature selection algorithm has the method called redundant prediction, it will select upon the extracted data of the patient characterization.

***MODULE NAME*** **:** Omics Identification

**BRIEF DESCRIPTION:**

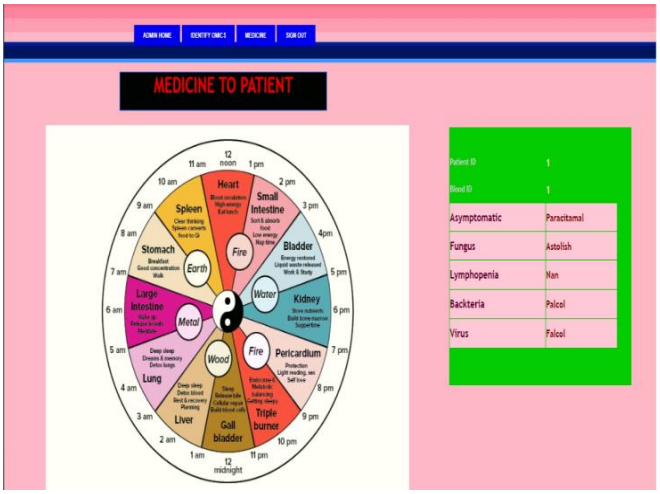
The omics identification will be done by feature extraction algorithm, in this the method used called relation prediction, and the molecules characterized are comparing with the omics characterization.

***MODULE NAME*** **:** Biomolecules Characterization

**BRIEF DESCRIPTION**

The biomolecules of each disease like, virus, fungus, bacteria are characterized in detailed manner. The higher will be updating the each disease characterization for identifying the biomarkers of each patient, to examine what kind of disease the caused by the patient.

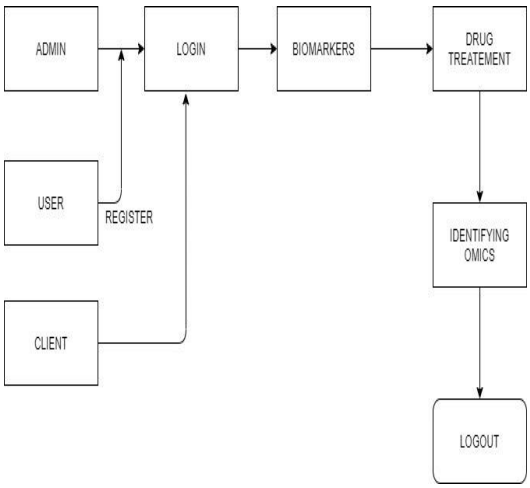
**OUTPUT**




**RESULT**

The higher will analyze the omics(structure of each disease like fungus, bacteria, and virus) from that the higher will report to the patient, from that they will proceed for the medicine.

**FLOWCHART**




**CONCLUSION**

The technology platform of genomics, proteomics and metabolomics ("-omic-" technologies) are high-throughput technologies. They increase substantially the number of proteins/genes that can be detected simultaneously and have the potential to relate complex mixtures to complex effects in the form of gene/protein expression profiles. By their nature, these technologies reveal unexpected properties of biological systems.

## REFERENCES

[1] F. S. Collins, and H. Varmus, "A new initiative on precision medicine," *N Engl J Med*, vol. 372, no. 9, pp. 793-5, Feb, 2015.

[2] R. Clifford, T. Louis, P. Robbe, and S. Ackroyd, "SAMHD1 is mutated recurrently in chronic lymphocytic leukemia and is involved in response to DNA damage," *Blood*, vol. 123, no. 7, pp. 1021-31, Feb, 2014.

[3] M. Rentoft, K. Lindell, P. Tran, A. L. Chabes, R. J. Buckland, D. L. Watt *et al.*, "Heterozygous colon cancer-associated mutations of SAMHD1 have functional significance," *Proc Natl Acad Sci U S A*, vol. 113, no. 17, pp. 4723-8, Apr, 2016.

[4] C. Schneider, T. Oellerich, H. M. Baldauf, S. M. Schwarz, D. Thomas, R. Flick *et al.*, "SAMHD1 is a biomarker for cytarabine response and a therapeutic target in acute myeloid leukemia," *Nat Med*, vol. 23, no. 2, pp. 250-255, Feb, 2017.

[5] S. Tan, Q. Gou, W. Pu, C. Guo, Y. Yang, K. Wu *et al.*, "Circular RNA F- circEA produced from EML4-ALK fusion gene as a novel liquid biopsy biomarker for non-small cell lung cancer," *Cell Res*, Apr, 2018.

[6] G. Biomarkers Definitions Working, "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework," *Clin Pharmacol Ther*, vol. 69, no. 3, pp. 89-95, Mar, 2001.

[7] F.-N. B. W. Group, *BEST (Biomarkers, EndpointS, and other Tools)Resource*, Silver Spring (MD), 2016.

[8] N. Cancer Genome Atlas Research, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nat Genet*, vol. 45, no. 10, pp. 1113-20, Oct, 2013.

[9] M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu *et al.*, "Proteomics. Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, pp. 1260419, Jan, 2015.

[10] S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate *et al.*, "COSMIC: somatic cancer genetics at high-resolution," *Nucleic Acids Res*, vol. 45, no. D1, pp. D777-D783, Jan, 2017.

[11] A. Mardinoglu, J. Boren, U. Smith, M. Uhlen, and J. Nielsen, "Systems biology in hepatology: approaches and applications," *Nat Rev Gastroenterol Hepatol*, vol. 15, no. 6, pp. 365-377, Jun, 2018.

[12] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biol*, vol. 18, no. 1, pp. 83, May, 2017.

.

Volume: 6, Issue: 3,Special Issue: 1 ,Mar,2021 ,ISSN_NO: 2321-3337