# IDENTIFYING MOLECULAR BIOMARKERS FOR DISEASES WITH MACHINE LEARNING BASED ON INTEGRATIVE OMICS

**Hanumanthkari Payal[1], Ramanpreet Kaur[2], Saba Khanum[3], Shruthi B[4]**

*Student, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India[1,2,3]*

*Assistant Professor, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India [4]*

---

**Abstract -** Molecular biomarkers are certain molecules or set of molecules that can be of help for diagnosis or prognosis of diseases or disorders. We present a comprehensive survey on the recent progress of identification of molecular biomarkers with machine learning approaches. Specifically, we categorize the biomarkers into diagnostic and prognostic using machine learning algorithms. In addition, we further find the survival analysis of the disease if it's prognostic and if it's diagnostic we find the prediction disease.

***Key Words*: *Molecular biomarker, machine learning, diagnosis, prognosis***

## 1. INTRODUCTION

Bioinformatics is an interdisciplinary field mainly involving molcular biology and genetics, computer science, mathematics, and statistics. Data intensive, large-scale biological problems are addressed from a computational point of view.

In recent years, increasing amounts of genomic data have become available. The size, type, and structure of these data have also been growing at an unprecedented rate. Gene expression, single nucleotide polymorphisms (SNP), copy number variation (CNV), proteomic, and protein-protein interactions are some examples of genomic and proteomic data produced using high throughput technologies such as microarrays, array comparative hybridization and mass spectrometry. Each of these distinct data types provides a different, partly independent and complementary view of the whole genome. However, elucidation of gene function and other aspects of the genome may require more information than is provided by one type of data. The amount and type of biological data are expected to increase even further (e.g., methylation, alternative splicing, transcriptomic, metabolomic, etc.). This proliferation of experimental data makes systematic integration an important component of genomics and bioinformatics. Data integration is increasingly becoming an essential tool to cope with the increasing amount of data, to cross-validate noisy data sets, and to gain broad interdisciplinary views of large genomic and proteomic data sets. Instances of combining and synthesizing data have increased considerably in the last several years and the need for improved and standardized methods has been recognized.

Technological advances allow biomedical researchers to collect a wide variety of omics data on a common set of samples. Data repositories such as The Cancer Genome Atlas (TCGA) provide multiple types of omics data, thus enabling in-depth investigation of molecular events at different stages of biology and for different tumor types. However, the latter task requires developing methods for data integration, a topic that has received increased attention in the literature. The classification of the approaches presented in the literature as multi-omics methods is a non-trivial task for at least three reasons. First, most of the computational approaches developed so far are pipelines of analysis that apply several methods to carry out a sequence of tasks; therefore, different pipelines share some methods: for example, partial least squares regression is included in both Integromics and sMBPLS.

Technologies used in sequencing of the human genome are dramatically reshaping the research and development pathways for drugs, vaccines, and diagnostics. The growth in the number of molecular entities entering the drug development pipeline has accelerated as a consequence of powerful discovery and screening technologies such as combinatorial chemistry, mass spectrometry, high throughput screening, cell- and tissue-based DNA microarrays, and proteomic approaches. As a consequence, there is an escalating number of therapeutic candidates, which has caused the need for new technologies and strategies to streamline the process to make safe and effective therapies available to patients.

A key challenge for integration methods is dealing with heterogeneous data. Data from different sources are difficult to compare due to inherent discrepancies. Different genomic variables are measured and collected in different ways, and they are associated with different types

of noise and confounding effects. Most importantly, they represent different aspects of the biological system. The discrepancy among data sources contributes to a useful multifaceted view of the system, but it also brings forth a new level of complexity that makes it hard to distinguish the coordinated signal.

The need for integration of heterogeneous data measured on the same individuals arises in a wide range of clinical applications as well.

There are many integration techniques that deal with the complexity of multiple sources by relying on prior knowledge of the relationships that connect them. Some procedures seek to map different experimental data types, such as gene expression (GE), miRNA expression (ME) and copy number variation to a common space of known biological pathways or sets. Others select features or assign weights to features based on prior knowledge, possibly using such information in a linear-based model. All these approaches require the consultation of an external resource, such as signaling pathways or gene interaction networks. While this supervised approach is convenient, it relies heavily on the external information being valid and representative, which is not always guaranteed, even in the modern era of data availability. In addition, relating variables based on previously established findings can introduce an element of bias and subjectivity that hinders the discovery of new associations.

Data integration is now a very commonly used notion in life sciences research. As of 2006 there were 1,062 papers explicitly mentioning "data integration" in their abstract or title, whereas this number has more than doubled in 2013 (2,365). However, there is still no unified definition of data integration, nor taxonomy for data-integration methodologies despite some recent efforts on this topic. In February 2013, the FP7 STATegra project and the COST Action SeqAhead, two EU-funded initiatives on the bioinformatics of high-throughput data, organized in the city of Barcelona the "Workshop of Omics and Data Integration", with the aim of reviewing current technologies on omics data production and the available methods for their integrative analysis. The workshop consisted of contributed talks, sessions for open discussion and we included an on-line survey to investigate the current opinions of the research community on this topic. Three major conclusions were extracted from the Barcelona workshop. First, there is a clear need for revisiting the concepts of data integration and stating available resources in this field; second, it was advantageous to extend our survey to a broader audience of scientists in life sciences, and third the commitment of organizers to publish the discussed topics, contributions and outcome of the public survey in a relevant journal is an important driver to spearhead further discussion in the community. In this supplement we discuss these three conclusions in some detail. In this introductory article we review current definitions of data integration and describes it formally as the combination of two challenges: data discovery and data exploitation. We briefly list major public efforts in creating resources (datasets, methods and workshops) for data integration. We also present the results of the extended community survey, which took place between February and March 2013 and on the basis of the survey we extract a couple of conclusions which warrant further elaboration in the community. We introduce the contributions of the papers collected in this supplement within the context of the discussed data integration topics and stated community needs.

The term data integration refers to the situation where, for a given system, multiple sources (and possible types) of data are available and we want to study them integrative to improve knowledge discovery. In the GLY example system we could have two datasets describing the system, one containing information about gene expression at the mRNA level and the other describing the CpG DNA methylation profile. In several studies where gene expression and DNA methylation data were available, the genome-wide relationships between DNA methylation and gene expression have been investigated in order to infer generic rules to questions such as: "Does DNA methylation regulation occurs at CpG islands and/or shores?", or "How does DNA methylation in promoters/gene-bodies/enhancers regulate gene expression?". These kinds of analyses have advanced our understanding of gene regulation by providing "generic rules yet with several exceptions" that associate epigenetic modifications with transcription. For instance, as a general rule CpG methylation in promoters in mammals was found to be anti-correlated with gene expression, while CpG methylation in gene bodies in mammals was positively correlated; yet these generic rules are observed as a trend, but are not necessarily true for all genes and/or for all biological situations.

To understand the challenges of data integration it is first required to define the term. The term "data integration" first appeared from the need to access different databases with overlapping content to provide "a redundancy free representation of information from a collection of data sources with overlapping content" which describes a need that appeared when the first databases were designed and it was required to connect several of them: "integration of multiple information systems aims at combining selected systems so that they form a unified new whole and give

users the illusion of interacting with one single information system". The aims of database integration were to make data more comprehensively available, and to increase the value of existing data by allowing previously difficult queries to be made upon it. Data mining is a major beneficiary from database integration. However this definition considers only access to data, and not exploitation of data, hence this definition of data integration is not fully applicable to life sciences research. We define data integration as the use of multiple sources of information (or data) to provide a better understanding of a system/situation/association/etc. Hence data integration, as defined here, is an action performed on a daily basis by most individuals, and a critical element in research.

## 1. LITERATURE SURVEY

Biomarkers serve a wide range of purposes in drug development, clinical trials, and therapeutic assessment strategies. Biomarkers can provide a basis for the selection of lead candidates for clinical trials, for contribution to the understanding of the pharmacology of candidates, and for characterization of the subtypes of disease for which a therapeutic intervention is most appropriate. Given this scenario, there are minimal public health consequences of an inaccurate reliance on a biomarker facilitated and strengthened by the use of appropriate biomarkers that measure biological parameters of disease and therapeutic response in humans. The realization of the potential benefits that surrogate endpoints can bring in expediting of the development of safe and effective therapies will require an increased understanding of the linkage of biomarkers to clinical endpoints and will necessitate high levels of scientific scrutiny and rigor.

On the basis of methodological aspects, we will consider two main criteria. The first is whether the approach uses graphs to model the interactions among variables. These approaches, designated as “network-based” (NB), take into account currently known (e.g. protein-protein interactions) or predicted (e.g. from correlation analysis) relationships between biological variables. In this class, graph measures (e.g. degree, connectivity, centrality) and graph algorithms (e.g. sub-network identification) are used to identify valuable biological information. Importantly, networks are used in the modeling of the cell’s intricate wiring diagram and suggest possible mechanisms of action at the basis of healthy and pathological phenotypes. Different types of sentiment analysis and its applications were presented. Algorithms and their originating references of various SA techniques are categorized and briefly explained.

The second criterion is whether the approach is bayesian (BY), that is, it uses a statistical model in which, starting from an a priori reasonable assumption about the data probability distribution, parametric or non-parametric, it is possible to compute the probability distribution making use of the Bayes’ rule; of course the posterior distribution depends on dataset measurements. In the network-based area, bayesian networks are another promising framework for the analysis multiomics data. Therefore, we will arrange integrative methods in four classes: network-free non-bayesian (NF-NBY), network-free bayesian (NF-BY), network-based non-bayesian (NB-NBY) and network-based bayesian (NB-BY) methods. We will give an overview of the methods that have been proposed for the analysis of at least two different types of omics datasets and describe with more details the specific mathematical grounds. In particular, we choose to consider in detail the mathematical aspects of the most common, representative or promising methods of each category.

With the rapid advances in biological high-throughput technology, generation of various kinds of genomic data is commonplace in almost every biomedical field. Effective data management and analytical approaches are essential to fully decipher the biological knowledge contained in the tremendous amount of experimental data. In the past decade, the accumulation of transcriptomic data mainly from microarray experiments was particularly significant, and resulted in several large public data depositories (such as Gene Expression Omnibus and Array Express). Similarly, genome-wide associate on studies (GWAS) are another example: thousands of GWAS have been performed world-wide and results and/or raw data for many are publicly available . It is common that multiple transcriptomic studies or GWAS are available for the same or related disease condition and each study has relatively small sample size with limited statistical power. Combining information from these studies to increase sensitivity and validate conclusions is a natural step. Such genomic information integration is akin to the classical meta-analysis in statistics where results of multiple studies of a similar research hypothesis are combined for a conclusive finding.

## 3. METHODOLOGY

### 3.1 Preprocessing

During preprocessing any redundancy from the dataset is removed. The dataset having any null value is also removed. For preprocessing of the data, feature scaling is used. Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. Standardization technique is used

to perform feature scaling. It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

### 3.2 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. 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 some cases, it is difficult to explain why and how the new features contribute to the good performance of the downstream learning algorithms.

### 3.3 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. The filter methods 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: ranks the features based on evaluation criteria and 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. 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.
NMF is a powerful tool for data reduction and exploration that has seen popular use in analyzing high-throughput genomic data. The method is related to PCA, except that it employs the constraint of non-negativity in lieu of orthogonality. As a result, NMF solutions are less uniquely defined but are more interpretable.
Given non-negative data matrix XN×M, NMF finds a non-negative factorization WH of rank D that best approximates X, typically in terms of the Frobenius norm.
$\min_{W,H} \|X - WH\|_{2F} \text{ s.t. } W \geq 0, H \geq 0.$

While Euclidean distance assumes a Gaussian distribution of values, alternative formulations of NMF using Bregman divergences have been proposed. Bregman divergences, which bear a strong connection with exponential families, encompass a wide range of distributional assumptions. Although we use Euclidean distance in the formulation of our method later, alternative loss functions may be accommodated via adjustments to the algorithm.

### 3.4 Classification

Classification is a process of categorizing a given set of data into classes. For classification we use K-Nearest Neighbors (KNN). K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures. It is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. 1990s standardization of hardware and software resultedin the ability to build modular systems. The increasing importance of software running on generic platforms hasenhanced the discipline of software engineering.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).
Classification is done by a majority vote to its neighbors. The data is assigned to the class which has the nearest neighbors. As you increase the number of nearest neighbors, the value of k, accuracy might increase.

### 3.5 Confusion Matrix

In Machine Learning (ML), you frame the problem, collect and clean the data, add some necessary feature variables (if any), train the model, measure its performance, improve it by using some cost function, and then it is ready to deploy.
A much better way to evaluate the performance of a classifier is to look at the confusion matrix. The general idea is to count the number of times instances of class A are classified as class B. Each row in a confusion matrix represents an actual class, while each column represents a predicted class. The confusion matrix gives you a lot of

information, but sometimes you may prefer a more concise metric.

Precision

precision = (TP) / (TP+FP)

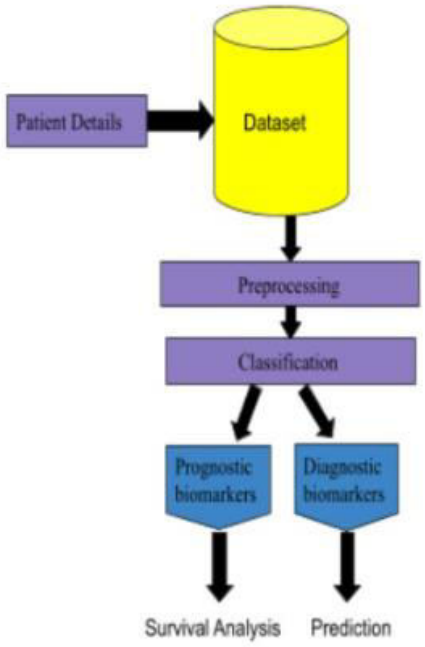
TP is the number of true positives, and FP is the number of false positives.

A trivial way to have perfect precision is to make one single positive prediction and ensure it is correct (precision = 1/1 = 100%). This would not be very useful since the classifier would ignore all but one positive instance.

Recall

recall = (TP) / (TP+FN)

## 4. ARCHITECTURE DIAGRAM




## 5. RESULT

The patient details such as patient Id, age, gender is considered. The biomarker database has been classified into prognostic and diagnostic biomarkers. After the biomarkers have been classified, the patient data is being used to tell whether the biomarkers found is prognostic or diagnostic.

If the biomarker is prognostic the patient details along with the medication and required message is displayed.

If the biomarker is diagnostic, patient details along with medication, possibly recommendation for the disease treatment, the doctor best suited for the treatment, hospital and the survival percentage is displayed.

## 6. CONCLUSIONS

Biomarkers can be the future of early prediction of diseases, thus making it possible for early treatment and cure of certain deadly diseases like cancer, Alzheimer's etc. Here the biomarkers are classified based on prognosis and diagnosis hence, the severity of the disease is known based on the classification. Here we present only certain disease such as cancer, kidney stone, etc. Further even more wide spectrum of diseases can be added and also their cure by combining convolutional neural network (CNN) and artificial neural network (ANN). Along with detection of the disease, we can predict the stage of the spread of the disease, the cure for the diseases and its chances of returning in the future (if any).

## ACKNOWLEDGEMENT

Sincere thanks to the college Atria Institute of Technology and also to dearest faculty of Information Science and Engineering Department for helping us throughout the completion of the project. The research team appreciates and heartily thanks project guide, our mentor Shruthi B ma'am for believing in us and in successful completion of the project.

## 8. REFERENCES

[1] X. Yu, G. Li, and L. Chen, "Prediction and early diagnosis of complex diseases by edge-network," Bioinformatics, Mar, 2014

[2] D.R.Hardoon,S.szedmak and J.shawe Taylor, "An overview with applications to learning methods", neural computation, Dec 2004

[3] Z.yang,and G.Michailidis,"A non negative matrix factorization methods for detecting modules in heterogeneous omics multi-modal data ",Jan 2016

[4] Neural networks ",May 2018 J.Zhang,W.peng and L.wang,"LeNup:Learning nucleosome positioning from DNA sequences with improved convolutional .

[5] S.Kim,D.Kang,Z.huo,"Meta-analytic principle component analysis in integrative omics applications ",April 2018

[6] Y.Su,T.M.Murali,V.pavlovic,"Identification of diagnostic gene based on expression data ",Aug 2003.

[7] Y.Li,A.Ngom,"A review on ML principles for multi view biological data integration ",Mar.

[8] D.Sun,X.Ren,P.Csermely,"Discovering cooperative Biomarkers for heterogeneous complex disease diagnosis",Jan 2019.

[9] Y.Nam,J.Cho,H.Shin,"Disease gene Identification based on generic and disease specific genome network",Jan 2019.