Bioinformatics Advancements for Detecting Epidemic Disease Using Machine Learning Approaches

Bikash Baruah and Manash Pratim Dutta

Abstract In the twentieth century, many researchers have started working on bioinformatics for disease biomarker detection using genetic information, i.e., DNA microarray dataset and RNA sequencing dataset with machine learning approaches. The journey of this concept starts with the classification technique on DNA microarray dataset by comparing it with reference genome or by deNovo (without reference genome) technique, and lots of different tools were published in different publications. Later, with the availability and advancement of computational power many researchers started working on large RNA sequencing dataset and some tools are published again with significant features. Nowadays, also this area is like a newborn baby and several challenges are still not solved, but it does not have a proper guideline for new researchers to face those challenges. After analyzing so many tools on DNA as well as RNA, we are able to summarize these works with a common workflow, and in this paper, we have proposed a generalized workflow for detecting epidemic diseases like HIV-AIDS, Cancer using machine learning approaches.

Keywords DNA microarray · RNA sequencing · Genome · Sanger · NGS · Differential co-expression

1 Introduction

A recent advancement of bioinformatics [1, 2] is in trend where machine learning approaches are used on DNA microarray [3–6] and RNA-seq dataset [7–10] to identify the progression of epidemic diseases. The effectiveness and reliability of this approach are far better than the traditional techniques. Researchers are working continuously to develop cost-effective and robust algorithms. To obtain the input

B. Baruah (✉) · M. P. Dutta
Department of Computer Science and Engineering, National Institute of Technology Arunachal Pradesh, Papum Pare, India
e-mail: bikash.phd@nitap.ac.in

M. P. Dutta
e-mail: manashpdutta@nitap.ac.in

P. K. Mallick et al. (eds.), *Electronic Systems and Intelligent Computing*, Lecture Notes in Electrical Engineering 686, https://doi.org/10.1007/978-981-15-7031-5_100

sequence, basically there are two sequencing methods, i.e., Sanger [11, 12] and NGS [13, 14], and these can be applied either in extracted DNA or RNA of any living being and result obtained will be microarray and RNA-seq, respectively. Now co-expression analysis [15–17], differential expression analysis [16], and differential co-expression analysis [18–22] or hybridized analysis (combination of these analyses) can be used through bi-clustering [23–25] or tri-clustering [26] techniques to detect the disease biomarker [27]. Here, we are proposing one generalized workflow which is fitted in almost all researches of this area. Further, we will extend our work to design robust and cost-effective algorithm to apply Cancer and HIV progression human dataset to identify the highly affected genes.

2 Proposed Model

We have proposed a model given in Fig. 1 which gives a complete workflow starting with DNA and RNA extraction from living cells or tissues followed by sequencing and co-expression analysis. In each step, we try to explain different available techniques. The workflow discussed in this model will certainly help the new researchers of this field, because it was never explained before in such a simple, systematic and step-by-step manner how wet lab and dry lab processes are combined together for detecting disease affected genes. Once sample extraction followed by sequencing is being completed in wet lab, the output of sequencing is taken as the input for dry lab for data analysis. In the following sections, different modules are explained.

2.1 Sample Extraction

Sample can be of two types: DNA and RNA and their extraction process from living or conserved cells, tissues, or virus particles are also different. Though, nowadays, many advanced kits are available for high-quality DNA and RNA extraction [28], the basic steps are almost similar to each other.

2.1.1 DNA Extraction Procedure

Step 1 Cell lysis to release the DNA.
Step 2 Centrifuge the sample to separate the DNA from other cellular debris and proteins.
Step 3 Use chilled isopropanol to precipitate the DNA.
Step 4 Wash DNA properly with ethanol.
Step 5 Gel electrophoresis for quality and quantity check of DNA.

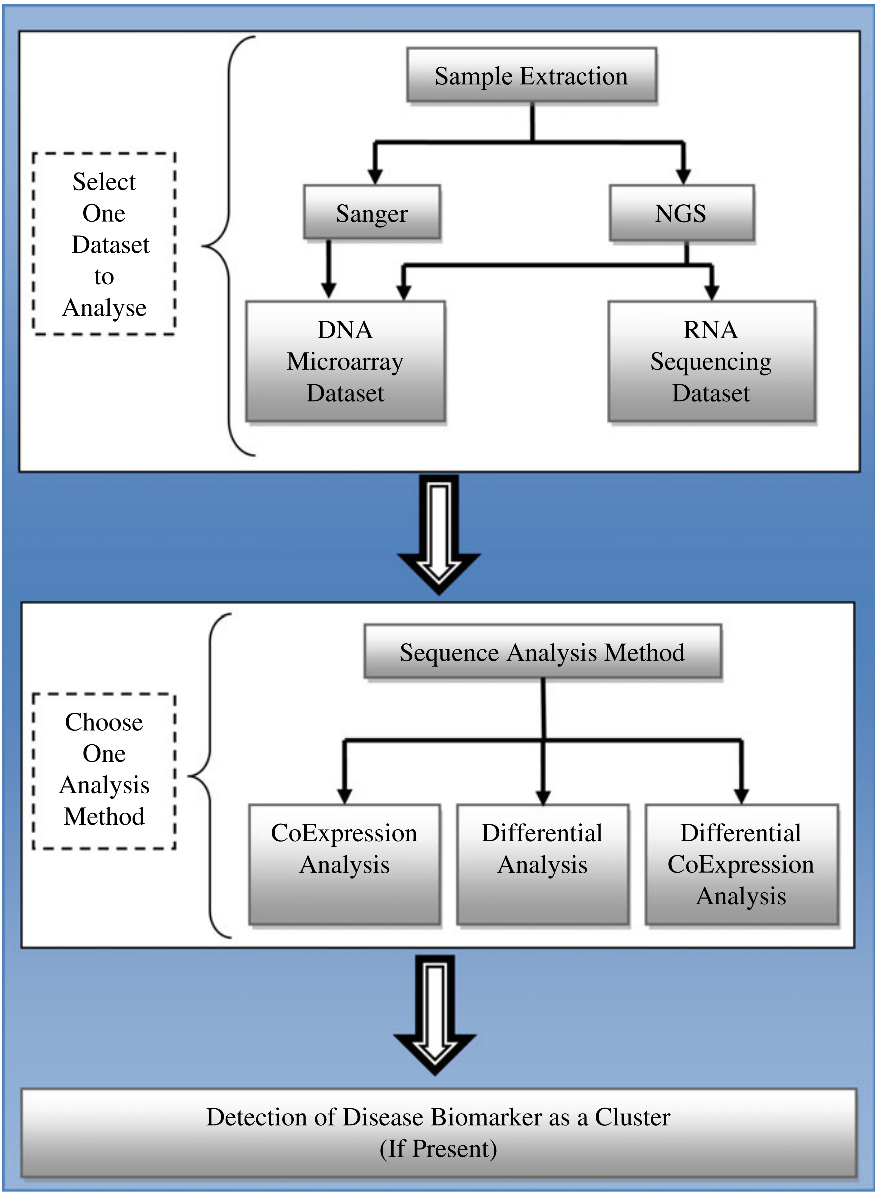
Select One Dataset to Analyse
Sample Extraction
Sanger
NGS
DNA Microarray Dataset
RNA Sequencing Dataset
Choose One Analysis Method
Sequence Analysis Method
CoExpression Analysis
Differential Analysis
Differential CoExpression Analysis
Detection of Disease Biomarker as a Cluster (If Present)

Fig. 1 Proposed model

2.1.2 RNA Extraction Procedure

Step 1 Cell lysis and dissolution.
Step 2 Denaturation of proteins and DNA.
Step 3 RNases inactivation.
Step 4 Removal of cellular components.
Step 5 Precipitation of RNA.
Step 6 Gel electrophoresis for quality and quantity check of RNA.

Once, high-quality sample is extracted from cells, tissues, or virus particles; it becomes ready for sequencing either by Sanger or NGS.

2.2 Sanger Sequencing

Frederick Sanger and his colleagues developed Sanger sequencing in 1977, which is known as the first generation sequencing. In complete Human Genome Project, Sanger sequencing is used and completed in 2003. The output of Sanger sequencing gives high-quality data with low noise and robustness. Sanger sequencing method uses dideoxynucleotides (ddNTP) with a hydrogen atom instead of $3^{'}$ hydroxyl group to sequence the deoxyribonucleic acid (DNA). These modified ddNTPs are able to terminate the polymerization of DNA. Here, DNA sample is divided into four separate samples and each of the four samples contains DNA polymerase and deoxynucleotides (dATP, dGTP, dCTP, and dTTP). In each sample, one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) is added. When the bases of unknown strand bind with its complementary ddNTPs, the polymerization will stop as ddNTPs and fragments of DNA are produced. Further, the samples from all four vessels are collected and determine their size by agarose gel electrophoresis. Unknown sequence is determined by arranging the size of fragments from lowest to highest in $5'$ to $3'$ order. Many copies of DNA fragments with different lengths are generated to compute the final DNA Microarray dataset which will be further analyzed by applying expression analyzing algorithm to detect the disease biomarker.

2.3 Next-Generation Sequencing

Next-generation sequencing (NGS), a more cost-effective and efficient sequencing technology compared to Sanger sequencing, is used to sequence both DNA and RNA. Different machines are available for next-generation sequencing (NGS), but basically it follows three steps to complete the process.

2.3.1 Sample Preparation

It needs custom adapter sequence either by ligation or amplification. These adapter sequences provide universal primers for library hybridization to the sequence chip.

2.3.2 Sequencing Through Machines

Each library fragment is amplified and attached with DNA linkers to hybridize the adapters. This creates clusters of DNA, and each cluster is an individual sequencing read.

2.3.3 Collect Output Data

At the end of sequencing, raw data in the form of reads will be available which can further be analyzed to retrieve the more meaningful and informative result.

NGS can be implemented for both DNA microarray and RNA sequencing. In microarray datasets, gene intensities are in normal distribution; whereas in RNA-seq, it follows either Poisson or negative binomial distribution. The major advantages of RNA-seq over microarray dataset are: DNA microarray has very less sensitivity to gene expressions compared to RNA-seq dataset. RNA-seq can measure approximately 70,000 non-coding [29] RNAs which have an important role in disease biomarker detection; but, it is not possible in microarray.

2.4 Co-expression Analysis

Co-expression analyses are done generally in three steps.

Firstly, individual relationships among genes have to be calculated based on mutual information on each pair of genes. These information are stored in a matrix to describe the similarity or co-expression among the expression patterns of different genes across all the samples. Let us consider an example of five gene co-expression matrix as shown in Table 1.

In Table 1, we can see that maximum and minimum values are one and zero, respectively, for completely identical and complete dissimilar gene pair. Different ways of correlation measures, i.e., (1) Spearman's or Pearson's correlations [30, 31], (2) least absolute error regression [32], and (3) Bayesian algorithm [33] can be used to derive the co-expression matrix shown in Table 1. The values of Table 1 are only to describe the pattern of a co-expression matrix. Bayesian algorithm and least absolute error regression have the advantage to identify causal links.

Secondly, co-expression network has to be constructed using genes as nodes and co-relation between the nodes as edges. Edge can be either weighted versus unweighted

Table 1 5 × 5 co-expression matrix

	Gene1	Gene2	Gene3	Gene4	Gene5
Gene1	**1**	0.50	0.67	00	0.32
Gene2	0.50	**1**	0.39	0.80	0.45
Gene3	0.67	0.39	**1**	0.20	0.17
Gene4	00	0.80	0.20	**1**	0.78
Gene5	0.32	0.45	0.17	0.78	**1**

and signed versus unsigned. Thickness of edge shows the weight of the edge, and the value lies between zero and one.

Weighted Versus Unweighted
Unweighted edged network is the simplest way of constructing co-expression network where interaction between node pairs is binary, i.e., either 0 or 1. By considering the correlation of all gene pairs or node pairs above, a certain threshold to be connected (i.e., 1) and all others be disconnected (i.e., 0).

In a weighted edged network, all nodes are connected to each other with a weighted edge consists of continuous values determining the co-relation between the nodes where value determines the strength. Weighted edge can be of two types: signed and unsigned edges.

Signed Versus Unsigned
In a signed network, edge correlation values lie between −1 (perfect negative correlation) and 1 (perfect positive correlation). An unsigned edge network assigns the correlation values between 0 and 1 so that values less than 0.5 indicate negative correlation and values greater than 0.5 indicate positive correlation.

Thirdly, co-expressed genes are clustered using bi-clustering or tri-clustering technique to group the genes with similar expression patterns across multiple samples. Some well-known clustering algorithms are K-means clustering [34], hierarchical clustering [34], THD-tricluster [26], shifting-and-scaling correlation clustering [35], etc. The clusters can be interrogated to identify regulators, functional enrichment, and hub genes for a potential disease gene by using guilt-by-association (GBA) [36] approach, while differential co-expression analysis gives the advantage of comparing modules in different conditions for better identifying disease regulators.

2.5 Differential Expression Analysis

Differential analysis has been done by comparing gene expression datasets of different conditions. For disease detection, minimum two samples have to be considered; one dataset of normal or healthy conditions and another in unhealthy or disease affected conditions. Different statistical tests like *t*-test, *z*-test, chi-square test are applied to analyze whether expressions are in up-regulation (disease is growing) or down-regulation (in control) states. In unhealthy conditions, if more number of samples with a fixed interval can be collected, then differential analysis gives more informative result. Bi-clustering and tri-clustering techniques are used for two datasets and more than two datasets, respectively, to group the genes showing responses in the same conditions.

2.6 Differential Co-expression Analysis

Differentially co-expressed analysis is to identify the patterns of correlated gene expression in different conditions. It will always give a more informative picture of the dynamic changes in the gene regulatory networks by comparing the transcriptome of same genome in two conditions. For example, one cluster of genes strongly correlated in one condition may no longer be strongly correlated in another condition. Hence, differential co-expression gives high response to potential disease adaptation in different environments. Differential co-expression analysis can be done in three ways:

2.6.1 Targeted Differential Co-expression

Differential co-expression analysis starts with targeted approach. In general, it is completed in three steps.

Firstly, pre-defined clusters are being surveyed with known annotation file to analyze in different conditions.

Secondly, correlation among genes of individual clusters as well as the correlation within group of clusters has been derived by using the gene correlation expression.

Finally, comparison is done between gene co-expression values in multiple environmental conditions.

2.6.2 Untargeted Differential Co-expression

Untargeted differential co-expression is the latest approach among all bioinformatics sequence analyzer. It is also done in three steps. Unlike targeted in the first step, correlated genes have to be detected which shows different significant behavior in different conditions. Once clustering is completed, rest two steps are similar with targeted approach. In 2009, Southworth et al. [5] applied this approach for the first time which is based on purely untargeted approach for detecting the mice genetic modules correlation with respect to age.

2.6.3 Semi-targeted Differential Co-expression

It is somewhat in between targeted and untargeted, where pre-defined clusters with partial annotation files are used. A strong disadvantage of semi-targeted approach is that it only concerns with those genes which emerge with clusters at least in anyone different environmental conditions.

2.7 Cluster Detection

Once the analysis is being completed by using any of the analysis methods, viz. traditional co-expression analysis, differential expression analysis, or differential co-expression analysis, the affected genes will be discovered. Its efficiency depends on the effectiveness of the algorithm designed by the researchers. Then, these genes are clustered as a module so that this cluster can be used in further drug design.

3 Conclusion

In this paper, we have tried to explain a workflow in a sequential manner for detecting epidemic diseases affected genes using different bioinformatics advancements. In future, we are going to implement these approaches on different DNA, RNA samples to detect Cancer and HIV-AIDS affected genes and will try to cluster them separately, so that our result can help the drug designers at genetic level.

References

1. Tomasz P, Szymon W, Jacek B (2016) Computer representations of bioinformatics models. Curr Bioinform 11(5):551–560
2. Agbachi CPE (2017) Pathways in bioinformatics: A window in computer science. Int J Comput Trends Technol 49(2):83–90
3. Sardaraz M, Tahir M, Ikram AA (2016) Advances in high throughput DNA sequence data compression. J Bioinf Comput Biol 14(3):18
4. Ge SX (2017) Exploratory bioinformatics investigation reveals importance of junk DNA in early embryo development. BMC Genom 18(1):200
5. Chen S, Liu M, Zhou Y (2018) Bioinformatics analysis for cell-free tumor DNA sequencing data. In: Computational Systems Biology. Humana Press, New York, NY, USA, pp 67–95
6. Zhang J, Huang K (2017) Pan-cancer analysis of frequent DNA come thylation patterns reveals consistent epigenetic landscape changes in multiple cancers. BMC Genom 18:1045
7. Van Dam S, Craig T, de Magalhaes JP (2015) Gene friends: a human RNA-seq-based gene and transcript co-expression database. Nucl Acids Res 43:1124–1132
8. Zeisel A, Munoz-Manchado AB, Codeluppi S et al (2015) Brain structure cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. 347:1138–1142

9. Fiannaca A, La Rosa M, La Paglia L et al (2015) Analysis of miRNA expression profiles in breast cancer using biclustering. BMC Bioinform 16
10. Xue Z, Huang K, Cai C et al (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature 500:593–597
11. Sanger F (1980) Google Scholar. http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1980/sanger-bio.html
12. Hutchison C (2007) DNA sequencing: bench to bedside and beyond Nucleic Acids. Nucl Acids Res. 35:6227–6237
13. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends genet. Trends Genet 24:133–141
14. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145
15. Glass K, Huttenhower C, Quackenbush J (2013) Passing messages between biological networks to refine predicted interactions. PLoS One 8
16. De Smet R, Marchal K (2010) Advantages and limitations of current network inference methods. Nat Rev Microbiol 8:717–729
17. Yue F, Cheng Y, Breschi A et al (2014) A comparative encyclopedia of DNA elements in the mouse genome. Nature 515:355–364
18. Amar D, Safer H (2013) Dissection of regulatory networks that are altered in disease via differential co-expression. PLoS Comput Biol 9
19. Zeisel A, Munoz-Manchado AB, Codeluppi S et al (2015) Brain structure cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347:1138–1142
20. Bhar A, Haubrock M, Mukhopadhyay A et al (2013) Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell. Algor Mol Biol 8
21. Fiannaca A, La Rosa M, La Paglia L (2015) Analysis of miRNA expression profiles in breast cancer using biclustering. BMC Bioinform 16
22. Kakati T, Bhattacharyya DK, Barah P, Kalita JK (2019) Comparison of methods or differential co-expression analysis for disease biomarker prediction. Comput Biol Med 10:100–103
23. Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform 1:24–45
24. Wang YK, Print CG, Crampin EJ (2013) Biclustering reveals breast cancer tumour subgroups with common clinical features and improves prediction of disease recurrence. BMC Genom 14:102
25. Oghabian A, Kilpinen S, Hautaniemi S, Czeizler E (2014) Biclustering methods: biological relevance and application in gene expression analysis. PloS One 9
26. Kakati T, Kashyap H, Bhattacharyya DK (2016) THD-module extractor: an application for CEN module extraction and interesting gene identification for Alzheimer's disease. Sci Rep 6
27. Kakati T, Bhattacharyya DK, Barah P, Kalita JK (2019) Comparison of methods for differential co-expression analysis for disease biomarker rediction. Comput Biol Med 10:113
28. Tan SC, Yiap BC (2009) DNA, RNA, and protein extraction: the past and the present. Hindawi Publ Corp J Biomed Biotechnol Article ID 574398
29. Zhao Y, Li H, Fang S et al (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucl Acids Res 44:203–208
30. Guttman M, Donaghey J, Carey BW et al (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477
31. Ala U, Piro RM, Grassi E et al (2008) Prediction of human disease genes by human-mouse conserved coexpression analysis. PLoS Comput Biol 4
32. van Someren EP, Vaes BL, Steegenga WT et al (2006) Least absolute regression network analysis of the murine osteoblast differentiation network. Bioinformatics 22:477–484
33. Friedman N, Linial M, Nachman I et al (2000) Using Bayesian networks to analyze expression data. J Comput Biol 7:601–620
34. Haeseleer PD (2005) How does gene expression clustering work? Nat Biotechnol 23:1499–1501
35. Ahmed H, Mahanta P, Bhattacharyya D, Kalita J (2014) Shifting-and-scaling correlation based biclustering algorithm. IEEE/ACM Trans Computat Biol Bioinform 11:1239–1252

36. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I (1998) The transcriptional program of sporulation in budding yeast. Science 282:699–705