ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Analysis of Microarray based Biological Pathway using Data Mining

N.Sevugapandi [1], C.P.Chandran [2]

Assistant Professor, Department of Computer Science, Sri Kaliswari College, Sivakasi. India[1]

Associate Professor, Department of Computer Science, Ayya Nadar Janaki Ammal College, Sivakasi, India[2]

ABSTRACT: Information technologies dominate almost all fields in the world, especially in the bio world. From the beginning, the collection of biological data has been increasing at explosive rates due to improvements of existing technologies and the introduction of new ones. The rapid progress of biotechnology and biological data analysis methods has led to the emergence and fast growth of a promising new field: bioinformatics. Biological pathways represent the biological reactions and interaction network in a cell. Each reaction is identified with its enzyme, which in turn is coded by certain gene(s). Diseases are increasingly identified with genetic modules, such as molecular pathways. Pathway-level analysis can provide important insights for making biological inferences and hypothesis from genetic data. Commonly the biological pathways are representing by diagrams and models its used in the analysis of genomic data and bridging the gap between diagrams and models allows not only the analysis of genomics data and interactions but also the visualisation of the results in a variety of different ways. Mining of bioinformatics data is an emerging area at the intersection between bioinformatics and data mining. In this paper we emphasise on pathway models, which can be used in research tools in bioinformatics.

KEYWORDS: Biological Pathway, Bioinformatics, Microarray

I. INTRODUCTION

Bioinformatics can be defined as the application of computer technology to the management of biological information. It is the science of storing, extracting, organizing, analyzing, interpreting and utilizing information from biological sequences and molecules. It has been mainly fueled by advances in DNA sequencing and mapping techniques. Some area of research in bioinformatics includes

- Sequence analysis
- Genome annotation
- Analysis of gene expression
- Analysis of protein expression
- Analysis of mutations in cancer
- Protein structure prediction
- Modeling biological systems

Recent day's pathway representations gained momentum as a research instrument. The places where, for example high-throughput genomics experiments, such as DNA microarrays, present the researcher with volumes of research data that are often too large for manual assessment. Mathematical methods like clustering or principal component analysis can be used to structure the large data volumes [2], but do not normally lead to increased understanding. Pathway diagrams can be used to present the outcome of such mathematical methods or genomics data directly. Biologically relevant changes are more visible when projected by pathway diagrams than when presented as large sets of tabular data.

This new role of the pathway representations in analysis creates specific requirements for their creation and curation. When a pathway diagram is used as an illustration, desktop publishing tools can be used to draw the diagram (Adobe Photoshop, Paintshop Pro, etc.). In this role, pathway entities and relations between entities have not to be made

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

explicit, as long as the diagram can be interpreted by human assessment. When pathway diagrams are used as a research tool, they should be available in a computer readable form. Not only every visible aspect of a pathway diagram needs to be made explicit, but also all relationships needed for analysis. For instance, visible genes products and metabolites need to be connected to a database entry that can be used to link them to experimental data and reactions. Reactions between metabolites or gene products need to be treated as edges in an interaction network to allow network analysis. Reactions between metabolites or gene products need to be treated as edges in an interaction network to allow network analysis.

II. MAPPING GENES TO PATHWAYS

The gene-pathway-disease model for understanding diseases mechanism requires first the identification of significant pathways from a set of genes. There are many algorithms that address this issue with different levels of sophistication. They can be grouped into four major types. One is based on over-representation of genes in certain pathways; another type is based on functional scoring of genes, which allows adjustments based on the correlation between phenotype and gene expression data; the third type also incorporates topological information of the pathways; the forth type – the gene set approach – focuses on using functional units comprised of a set of correlated genes to make pathway predictions.

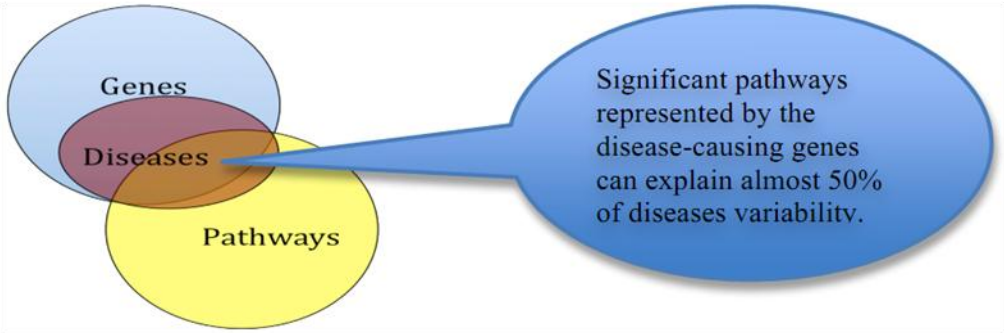

Fig 1. Model of Pathway Analysis

1) Over-representation

This approach starts with a set of differently expressed genes and uses statistical analysis to identify the Gene Ontology terms that are over-represented in the list of genes. This approach is limited in their accuracy because they do not incorporate "known interdependencies among genes in a pathway that can increase the detection signal for pathway relevance." [5]. In addition, they treat all genes identified from the gene expression data equally, which is not necessarily true. Thus, this approach can "produce many false positives when only a single gene is highly altered in a small pathway." [5]

2) Functional score of genes

This type of approaches incorporates functional indicators of the genes identified in a microarray experiment. An example of this is the Gene Set Enrichment Analysis (GSEA). It "ranks all genes based on the correlation between their expression and the given phenotypes, and calculates a score that reflects the degree to which a given pathway is represented at the extremes of the entire ranked list." [6]

Many experiments use different cell lines or different conditions for a certain disease phenotype. Genes may vary in stability depending on the cell type or disease being studied. [7] Thus, one of the drawbacks of the functional score approach is that the correlation with phenotypes could be compromised due to the variance of genes expression in different tissues or cell types. More details on this will be covered in a later section.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

3) Pathway topology

Pathway topology depicts the genetic interaction, indicating which genes are upstream of other genes, hence are prerequisites for the activation of the downstream genes. Changes in the expression level of the downstream genes may not affect the pathway as much as the upstream genes. Hence, such information is important to be incorporated in identifying the significantly enriched pathways.

4) Gene Set Approach

The methods abovementioned aim at inferring pathway representation directly from the gene expression data of each gene. All these approaches implicitly assume each gene as target for enrichment. The gene set approach, on the other hand, treats the known functionally related genes together as a group in calculating statistical significance. The null hypothesis is that genes of the same pathway are not co-regulated more strongly than a randomly selected group of genes without any functional relationship. This is more powerful than the previous approaches because the joint score of gene sets that are known to be in a functional relationship will increase the potential to detect subtle signals in gene expression data. [8] A number of methods have been proposed that work on the level of gene sets.

5. The Process of Pathway Curation

The application of pathway diagrams for data analysis requires the integration of knowledge from a wide variety of sources. Typically, a pathway diagram is created by integrating knowledge from biological databases, the scientific literature and intrinsic knowledge from domain experts [1]. Each of these three sources presents with specific challenges to extract and integrate the knowledge into a pathway diagram. Part of this challenge is the exponentially growing amount of available information that is scattered over a wide variety of sources and knowledge domains. Pathguide a website that lists pathway resources, (http://www.pathguide.org), already lists 261 different databases (October 2008) containing pathway knowledge [3]. As Cary [4] points out, there is the need to be aware of potential biases when integrating data from biological databases. Biological databases are often constructed around a specific species or with a specific set of research questions in mind.

III. DATA MINING AND BIOINFORMATICS

Many data mining techniques have been proposed to deal with the identification of specific DNA sequences. The most common include neural networks, Bayesian classifiers, decision trees, and Support Vector Machines (SVMs) [9][10]. Sequence recognition algorithms exhibit performance tradeoffs between increasing sensitivity (ability to detect true positives) and decreasing selectivity (ability to exclude false positives) [12]. However, as Li et al.[11] state, traditional data mining techniques cannot be directly applied to this type of recognition problems. Thus, there is the need to adapt the existing techniques to this kind of problems. Attempts to overcome this problem have been made using feature generation and feature selection [11][13]. Another data mining application in genomic level is the use of clustering algorithms to group structurally related DNA sequences.

1) Gene Expression Data Mining

The main types of microarray data analysis include (Piatetsky-Shapiro & Tamayo, 2003): gene selection, clustering, and classification. Piatetsky-Shapiro and Tamayo (2003) present one great challenge that data mining practitioners have to deal with. Microarray datasets -in contrast with other application domains- contain a small number of records (less than a hundred), while the number of fields (genes), and is typically in thousands. The same case is in SAGE data. This increases the likelihood of finding "false positives".

An important issue in data analysis is feature selection. In gene expression analysis the features are the genes. Gene selection is a process of finding the genes most strongly related to a particular class. One benefit provided by this process is the reduction of the foresaid dimensionality of dataset. Moreover, a large number of genes are irrelevant when classification is applied. The danger of overshadowing the contribution of relevant genes is reduced when gene selection is applied

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Clustering is the far most used method in gene expression analysis. Tibshirani et al [15] and Aas [16] provide a classification of clustering methods in two categories: one-way clustering and two-way clustering. Methods of the first category are used to group either genes with similar behavior or samples with similar gene expressions. Two-way clustering methods are used to simultaneously cluster genes and samples. Hierarchical clustering is currently the most frequently applied method in gene expression analysis. An important issue concerning the application of clustering methods in microarray data is the assessment of cluster quality. Many techniques such as bootstrap, repeated measurements, mixture model-based approaches, sub-sampling and others have been proposed to deal with the cluster reliability assessment

IV. DATA MINING TECHNIQUES IN BIOINFORMATICS

The entire human genome, the complete set of genetic information within each human cell has now been determined. Understanding these genetic instructions promises to allow scientists to better understand the nature of diseases and their cures, to identify the mechanisms underlying biological processes such as growth and ageing and to clearly track our evolution and its relationship with other species. The key obstacle lying between investigators and the knowledge they seek is the sheer volume of data available (fig 2). This is evident from the following figure 2 which shows the rapid increase in the number of base pairs and DNA sequences in the repository of GenBank.

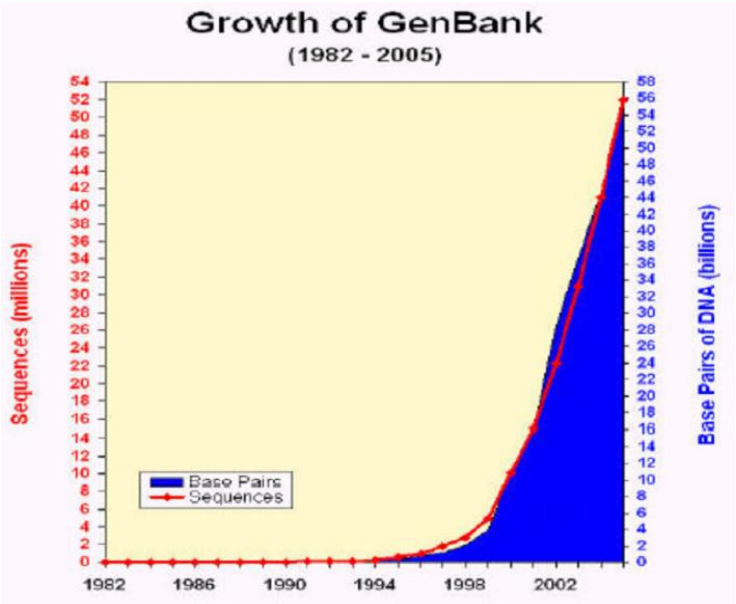

Fig.2 Growth of DNA Sequence

Biologists, like most natural scientists, are trained primarily to gather new information. Until recently, biology lacked the tools to analyze massive repositories of information such as the human genome database. Luckily, the discipline of computer science has been developing methods and approaches well suited to help biologists manage and analyze the incredible amounts of data that promise to profoundly improve the human condition. Data mining is one such technology. Here the list of some data mining techniques used in bio-data analysis.

- Association (A priori algorithm and Partition algorithm)
- Clustering (K-medoids algorithm and K-means algorithm)
- Classification using Decision tree(Classification and Regression Technique (CART))

1) Semantic Integration of Heterogeneous Data

One of the many complex aspects in biomedical data mining is semantic integration. Semantic integration combines multiple sources into a coherent data store and involves finding semantically equivalent real-world entities from several biomedical sources to be matched up. The problem arises when, for instance, the same entities do not have identical

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

labels, such as, gene id and g id, or are time asynchronous, as in the case of the same gene being analyzed at multiple developmental stages. There is a theoretical foundation for approaching this problem by using correlation analysis in a general case. Nonetheless, semantic integration of biomedical data is still an open problem due to the complexity of the studied matter (bioontology) and the heterogeneous distributed nature of the recorded high-dimensional data. Currently, there are in general two approaches:

- Construction of integrated bio-data warehouses or bio-databases and
- Construction of a federation of heterogeneous distributed bio-databases so that query processing or search can be performed in multiple heterogeneous bio-databases.

V. MICROARRAY

Molecular Biology research evolves through the development of the technologies used for carrying them out. It is not possible to research on a large number of genes using traditional methods. One can analyze the expression of many genes in a single reaction quickly and in an efficient manner using microarrays. Microarray technology has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning of the human body.

DNA microarrays [14] are created by robotic machines that arrange minuscule amounts of hundreds or thousands of gene sequences on a single microscope slide. Researchers have a database of over 40,000 gene sequences that they can use for this purpose. When a gene is activated, cellular machinery begins to copy certain segments of that gene. The mRNA produced by the cells bind to the original portion of the DNA strand from which it was copied. To determine which genes are turned on and which are turned off in a given cell, a researcher must first collect the mRNA molecules present in that cell. The researcher then labels each mRNA molecule by attaching a fluorescent dye. Next, the researcher places the labeled mRNA onto a DNA microarray slide(fig 3). The messenger RNA that was present in the cell will then hybridize or bind to its complementary DNA on the microarray, leaving its fluorescent tag. A researcher must then use a special scanner to measure the fluorescent areas on the microarray. Researchers frequently use this technique to examine the activity of various genes at different times. This activity is referred to as the gene expression value of the genes.

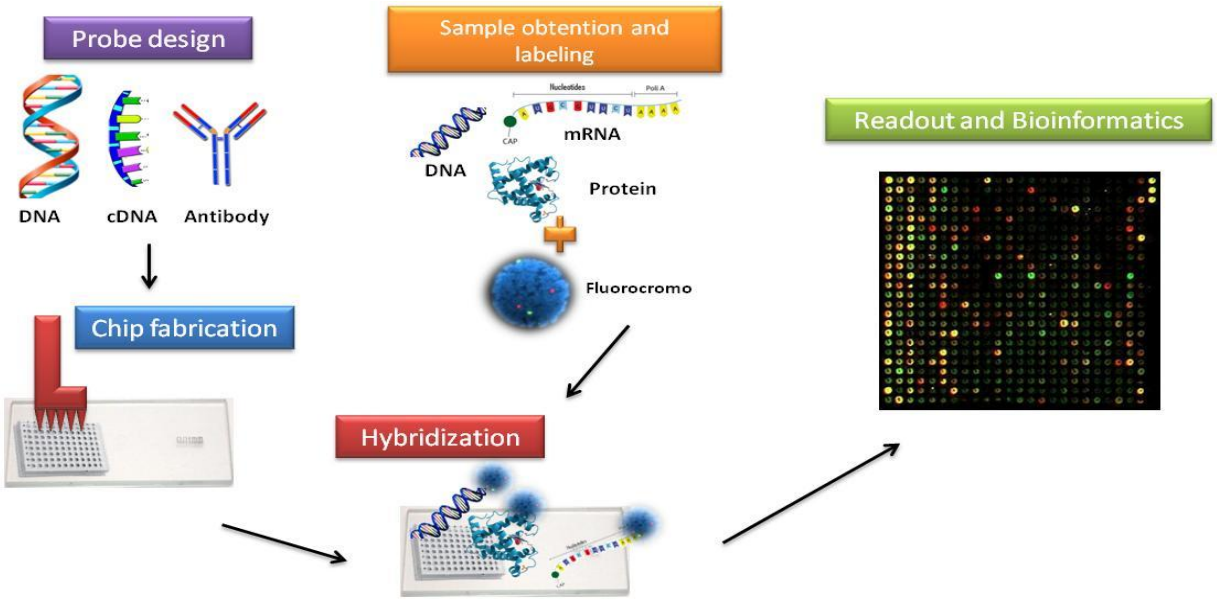

Fig.3 Example of Microarray usage

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

VI. CONCLUSION

This paper argues about bioinformatics, data mining techniques and need of biological pathway in detail. This study highlights data collection and analysis on results from genomic studies requires accurate and complete pathway models and the corresponding diagrams need to be interpretable by the researcher. The paper explores multiple techniques to draw a gene sequences in pathway representation. The microarray concepts were discussed with the example experimental result. In future the author moves towards finding the novel method for generating biological pathway and gene simulation based on bioinformatics data.

REFERENCES

[1] Adriaens ME, Jaillard M, Waagmeester A, Coort SL, Pico AR, Evelo CT, *The public road to high-quality curated biological pathways*. Drug Discov Today 13:856–862, 2008.
[2] Allison DB, Cui X, Page GP, Sabripour M, *Microarray data analysis: from disarray to consolidation and consensus*. Nat Rev Genet 7:55–65, 2006.
[3] Bader GD, Cary MP, Sander C, *Pathguide: a pathway resource list*. Nucleic Acids Res 34:D504–D506, 2006.
[4] Cary M, Bader G, Sander C, *Pathway information for systems biology*. FEBS Lett 579:1815–1820, 2005.
[5] Vaske C., et al., *Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM*. Bioinformatics 26: 237-245, 2010.
[6] Draghici S., et al., *A systems biology approach for pathway level analysis.* Genome Res. 17: 1537-1545, 2007.
[7] Lee S, Jo M, Lee J, Koh SS, Kim S, *Identification of novel universal housekeeping genes by statistical analysis of microarray data.* J Biochem Mol Biol. Mar 31;40(2):226-31, 2007.
[8] Rahnenf¨uhrer J., Domingues F., Maydt J., Lengauer T. *Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data.* Statistical Applications in Genetics and Molecular Biology Vol. 3 Issue 1 Article 16, 2004.
[9] Ma, Q. and Wang, J.T.L., *Biological Data Mining Using Bayesian Neural Networks: A Case Study*. International Journal on Artificial Intelligence Tools, Special Issue on Biocomputing, 8(4), 433-451, 1999.
[10] Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T. and Muller, R.-K. *Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites. Bioinformatics*, 16(9), 799-807. 2000.
[11] Li, J., Ng, K.-S. and Wong, L. *Bioinformatics Adventures in Database Research.* Proceedings of the 9th International Conference on Database Theory, 31-46, Siena, Italy, 2003.
[12] Houle, J.L., Cadigan, W., Henry, S., Pinnamaneni, A. and Lundahl, S. *Database Mining in the Human Genome Initiative.* Whitepaper, Bio-databases.com, Amita Corporation. Available: http://www.biodatabases.com/ whitepaper.html. 2004.
[13] Zeng, F., Yap C.H.R. and Wong, L. *Using Feature Generation and Feature Selection for Accurate Prediction of Translation Initiation Sites*. Genome Informatics, 13, 192-200, 2002.
[14] Piatetsky-Shapiro, G. and Tamayo, P : *Microarray Data Mining: Facing the Challenges.* SIGKDD Explorations, 5(2), pp 1-5, 2003
[15] Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., & Brown, P. *Clustering methods for the analysis of DNA microarray data (Tech. Rep.).* Department of Statistics, Stanford University, Stanford, California, USA, 1999.
[16] Aas, K. *Microarray Data Mining: A Survey*. NR Note, SAMBA, Norwegian Computing Center.2001.