# ANALYSIS OF PROTEIN MICROARRAY DATA USING DATA MINING

N.Sevugapand[1] and Ms.M.Hemalatha[2]

[1]Research Scholar, Ph.D Part-Time, Category –B, Research and Development Center, Bharathiar University, Coimbatore, Tamilnadu
[2] Research Scholar, M.Phil, Department of Computer Science, Madurai Kamaraj University, Madurai.

## ABSTRACT

*Latest progress in biology, medical science, bioinformatics, and biotechnology has become important and tremendous amounts of biodata that demands in-depth analysis. On the other hand, recent progress in data mining research has led to the development of numerous efficient and scalable methods for mining interesting patterns in large databases. This paper bridge the two fields, data mining and bioinformatics for successful mining of biological data. Microarrays constitute a new platform which allows the discovery and characterization of proteins.*

## KEYWORDS

*Pathway, Microarray, Protein Microarray, Cluster & KEGG.*

## 1. INTRODUCTION

Bioinformatics is the interdisciplinary science of interpreting biological data using information technology and computer science. It deals with algorithms, databases and information systems, web technologies, artificial intelligence, software engineering, data mining, image processing, etc. Microarray technology has become crucial tool for large scale and high-throughput biology. It allows fast, easy and parallel detection of thousands of addressable elements in a single experiment. Protein microarray technology has been shown to be a useful tool for multiplexed detection and proteomics studies. In protein microarray, proteins are prepared, arrayed and analyzed at high spatial density, to be particularly powerful for analyzing gene function, regulation and a variety of other applications. Proteins are more challenging to prepare for the microarray format than DNA, and protein functionality is often dependent on the state of proteins, such as post-translational modifications, partnership with other proteins, protein sub cellular localization, and reversible covalent modifications. This paper presents an idea about pathways analysis of protein microarray data and describes the applications and technologies used to analyze protein microarray data, gives introduction about the tools and databases have been developed for pathway analysis.

## 2. ANALYSIS OF PATHWAY

### 2.1 Pathway Analysis

- Biological processes in a cell form complex networks among gene products. Pathway analysis tries to build, model, and visualize these networks. Pathway tools are usually associated with a database to store the information about biochemical reactions, the molecules involved, and the genes. Several tools and databases have been developed and

DOI:10.5121/ijbb.2016.6202

are widely used, including KEGG [2] (Kyoto Encyclopedia of Genes and Genomes) database (the largest collection of metabolic pathway graphs)

- EcoCyc/MetaCyc (a visualization and database tool for building and viewing metabolic pathways),
- GenMAPP (a pathway building tool designed especially for working with microarray data).

With the latest developments in functional genomics and proteomics, pathway tools will become more and more valuable for understanding the biological processes at the system level .

## 3. MICROARRAY ANALYSIS

### 3.1 Microarray databases

The diversity of microarray formats and types of experiments has made it difficult that any database format has imposed and no database system has emerged as the gold standard. Indeed there has been some agreement on the minimum information about a microarray experiment that needs to be stored (the MIAME standard is an acronym for this).

One can distinguish two levels at which databases systems have been developed.

1. Local database systems: The analysis of microarray data goes through a series of steps where different types of data, images, binaries, and text have to be processed. It requires to store them in an easily accessible way. Some systems such as BASE or caArray are powerful solutions for storing data and experiments but their use is far from being so extended as that of analysis software tools.
2. Public array repositories: The biological community has agreed, from the beginning of microarrays, that data from published experiments should be made publicly available. This has created the need for public microarray repositories where any user could store their data in a suitable form. At the same time it has made an impressive quantity of data available for re-analysis by anyone who wishes to do it, offering an unparalleled wealth of opportunities whose power is just starting to show.

### 3.2 Technology used to analyze Microarray

Microarray technology allows biologists to monitor genome-wide patterns of gene expression in a high-throughput fashion. Applications of microarrays have resulted in generating large volumes of gene expression data with several levels of experimental data complexity. For example, a “simple” experiment involving a 10,000-gene microarray with samples collected at five time points for five treatments with three replicates can create a data set with 0.75 million data points. Historically, hierarchical clustering was the first clustering method applied to the problem of finding similar gene expression patterns in microarray data. Since then many different clustering methods have been used, such as k-means, a self-organizing map, a support vector machine, association rules, and neural networks. Several commercial software packages, e.g., GeneSpring or Spotfire, offer the algorithms for microarray analysis. Today, microarray analysis is far beyond clustering. By incorporating a priori biological knowledge, microarray analysis can become a powerful method for modeling a biological system at the molecular level. For example, combining sequence analysis methods, one can identify common promoter motifs from the clusters of coexpressed genes in microarray data using various clustering methods.

### 3.3 Cluster analysis of microarray data

Clustering is a method that is long used in phylogenetic research and has been adopted to microarray analysis. A clustering algorithm is widely used because of its simple implementation. Alon et al [15] proposed Clustering algorithms have proved useful to help group together genes with similar functions based on gene expression patterns under various conditions or across different tissue samples. Zantema et al [16] described an algorithmic improvement for detection of frequently occurring patterns as well as modules in biological networks. They proved that this improvement is based on fact that finding frequent sub-network problem is reducible to the problem of finding maximal frequent item sets. They performed their experiments in metabolic pathways obtained from the KEGG [2] database.

The traditional algorithms for clustering are:

1. Hierarchical clustering.
2. K-means clustering.
3. Self-Organizing-Maps (SOM)

## 4. POTEIN MICROARRAY

Chandra et al., 2011 [17] proposed novel detection platform in protein microarray. Which described Protein microarrays are powerful tools to capture and measure proteins from biospecimen. A protein microarray typically consists of a small piece of plastic or glass coated with thousands of capture reagents. This technology could isolate and study many potential biomarker proteins.

Protein microarrays, an emerging class of proteomic technologies, are fast becoming critical tools in biochemistry and molecular biology. Two classes of protein microarrays are currently available: analytical and functional protein microarrays. Analytical protein microarrays, mostly antibody microarrays, have become one of the most powerful multiplexed detection technologies. Functional protein microarrays are being increasingly applied to many areas of biological discovery, including studies of protein interaction, biochemical activity, and immune responses. Great progress has been achieved in both classes of protein microarrays in terms of sensitivity, specificity, and expanded application.

### 4.1.1 Protein Microarrays Analysis

According to their applications, the planar protein microarrays have been classified in three main categories: analytical microarrays, reverse phase arrays (RPA) and functional microarrays [5,10]. On the other hand, microspheres bead based systems should also be considered, which use different size or color beads as a support of the capture agent to analyze the sample. In such microarray format, flow cytometry is coupled in order to support the identification of each specific binding according to the size, color and mean fluorescence intensity of conjugated fluorochromes [10].

### 4.1.2 Analytical Microarrays

This kind of microarray is used to determine parameters such as the binding affinity and specificity and to study protein expression levels in complex mixtures [11], but also they cover clinical applications such as studies in immunology or biomarkers detection [7] and they can be used to monitor differential expression profiles, such as protein patterns in response to environmental stress or differences among a healthy tissue and with respect to a pathological sample [11]. In addition, analytical microarrays imply direct labeling protocols of thousands of proteins, which might be another critical limitation. The chemical labeling of proteins can destroy epitopes by covalent combination of dyes or haptens. Moreover, only selected target proteins can be analyzed by antibody microarrays [10].

### 4.1.3 Reverse Phase Arrays

There are currently a variety of methods for microarray analysis. Reverse phase protein microarray analysis requires consideration of 1) spot placement on the array 2) background intensity and 3) a sufficient number of data points. In this case, cellular or tissue lysate or even serum samples are immobilized on the microarray surface and the detection is completed through an antibody against the target proteins. To achieve a higher fluorescent signal to be detected, a secondary antibody conjugated with a fluorochrome is added to the first one. This ensures the signal intensity is directly related with the specificity, the binding affinity and the accessibility of the antibody against the target protein [10]. The production of a functional map for cell signaling pathways from individual cells or tissues by RPA arrays has increased the interest on this kind of arrays with the objective of developing personalized therapies [5,7]. The proteins involved in RPA do not require labeling and only a little amount of protein is needed to produce the microarrays. However, fewer analyses can be analyzed due to the limited number of labeled antibodies for detection and also low availability of specific protein antibodies suitable for RPAs [5,8].

### 4.1.4 Functional Microarrays

Functional microarrays are composed of full length functional proteins or protein domains and study the biochemical characteristics and functions of native proteins, as well as peptides or domains highly purified through cell-based methods or by cell-free expression on the microarray [7]. They allow studying the whole proteome in a single assay. Functional microarrays are also employed to examine the diverse protein interactions: protein-protein, protein-DNA, protein-RNA, protein-phospholipids and protein-small molecules [11].

Cell-free based protein microarrays have been applied to immunological studies , vaccine development [12], early detection of biomarkers, biochemical activity protein-protein interaction studies [8], such as protein-protein, protein-DNA, protein-RNA, protein-phospholipids, and protein-small molecule interactions [11] and toxin detection [13]. Over the last few years, several in situ expressed microarrays have been developed such as: Protein in situ arrays (PISA), printing protein arrays from DNA (DAPA) arrays and Nucleic Acids Programmable Protein Arrays (NAPPA).

NAPPA is one of the most relevant microarrays in this field. The DNA templates are bound onto the slide surface; the protein of interest is encoded and a GST tag is added. This is a fusion protein with a tag, which will allow binding to the slide. The biotinylated DNA plasmid is attached through an avidin to the aminopropyltriethoxysilane (APTES)-coated surface. In addition, RRL is used to carry out protein expression. There are also anti-GST antibodies attached to the slide, where the fusion protein joins. As a result, an array with the expressed protein and its corresponding DNA is achieved all on the same slide [6].

NAPPA is a good cost-effective technique because of the small volume of cell-free extract required for protein expression. Also, the use of immobilized DNA allows storage of the array for a long time until the next procedure. The main drawback is the invested time to generate the cDNA with the protein of interest and the tag, but even this method does not achieve a pure protein. On the other hand, high yields of high quality DNA were obtained for immobilization by using a diamine-derivatized resin. It was also found that BSA improved the binding efficiency of DNA and that is why a master mix of cDNA, antibody, BS3 and BSA is used [6].

### 4.1.5 Challenges in Protein microarray

Protein microarray faces many Challenges which include:

1) Finding a surface and a method of attachment that allows the proteins to maintain their secondary or tertiary structure and thus their biological activity and their interactions with other molecules.
2) Producing an array with a long shelf life so that the proteins on the chip do not denature over a short time.
3) Identifying and isolating antibodies or other capture molecules against every protein in the human genome.
4) Quantifying the levels of bound protein while assuring sensitivity and avoiding background noise.
5) Extracting the detected protein from the chip in order to further analyze it.
6) The capacity of the chip must be sufficient to allow as complete a representation of the proteome.

## 5. Microarray Data Analysis Methods

The main types of data analysis needed to for biomedical applications include:

- Gene Selection – in data mining terms this is a process of attribute selection, which finds the genes most strongly related to a particular class.
- Classification – classifying diseases or predicting outcomes based on gene expression patterns, and perhaps even identifying the best treatment for given genetic signature
- Clustering – finding new biological classes or refining existing ones.

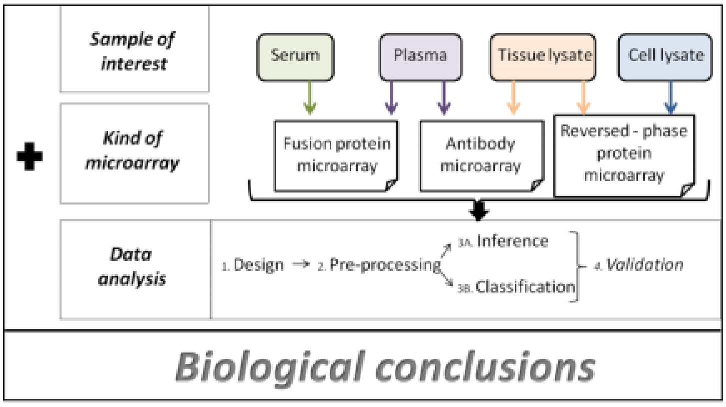

Figure 1. Workflow of protein microarray development

The wealth of information generated by protein microarrays may provide solid evidence concerning protein functions, their interactions and even their involvement in signaling pathways. Interestingly, these data can also be applicable as a tool for clinical diagnostics. Nevertheless, the translation of data into meaningful information requires automated data processing and handling. (Fig.1) Data processing and analysis is inherent to protein arrays. Thus, it becomes a crucial step in the search for solid biological conclusions. There are several strategies to analyze protein data, some of which have their origin in DNA microarray analysis, such as spot-finding on slide images, Z-score calculations and significance analysis of microarrays (SAM). However, concentration dependent analysis (CDA) has been specifically developed for protein microarrays [14].

## 6. CONCLUSIONS

Bioinformatics and data mining are developing as interdisciplinary science. Protein microarrays, an emerging class of proteomic technologies, are fast becoming critical tools in biochemistry and molecular biology. This paper discussed two classes of protein microarrays which are currently available: analytical and functional protein microarrays. This study has highlighted a various data analyze methods used in microarray, the examples presented here may give readers to get clear idea about protein microarray analysis. And also highlights the study papers on microarray data. Future works include research on advanced protein microarray analysis and also the research can incorporate the types of microarray such as DNA microarrays, MMChips for surveillance of microRNA populations, Peptide microarrays for detailed analyses or optimization of protein-protein interactions and Tissue microarrays. These enhancements can impact sensitivity and reproducibility.

## REFERENCES

[1] C. S. Patil, R.R.Karhe , M. A. Aher 2012. Development of Mobile Technology: A Survey. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 1, Issue 5.

[2] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acid Research, 28(1):27–30, 2000.

[3] G. Michal. Biochemical Pathways (Poster). Boehringer Mannheim, Penzberg, 1993.

[4] V. N. Reddy, M. L. Mavrovouniotis, and M. N. Liebman. Petri net representations of metabolic pathways. Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology (ISMB '93), pages 328–336, 1993.

[5] Chandra, H.; Reddy, P.J.; Srivastava, S. Protein microarrays and novel detection platforms. Expert Rev. Proteomics 2011, 8, 61–79.

[6] Chandra, H.; Srivastava, S. Cell-free synthesis-based protein microarrays and their applications. Proteomics 2010, 10, 717–730.

[7] Gonzalez-Gonzalez, M.; Jara-Acevedo, R.; Matarraz, S.; Jara-Acevedo, M.; Paradinas, S.; Sayagues, J.M.; Orfao, A.; Fuentes, M. Nanotechniques in proteomics: Protein microarrays and novel detection platforms. Eur. J. Pharm. Sci. 2012, 45, 499–506.

[8] Hultschig, C.; Kreutzberger, J.; Seitz, H.; Konthur, Z.; Bussow, K.; Lehrach, H. Recent advances of protein microarrays. Curr. Opin. Chem. Biol. 2006, 10, 4–10.

[9] Dasilva, N.; Diez, P.; Matarraz, S.; Gonzalez-Gonzalez, M.; Paradinas, S.; Orfao, A.; Fuentes, M. Biomarker discovery by novel sensors based on nanoproteomics approaches. Sensors 2012, 12, 2284–2308.

[10] Poetz, O.; Schwenk, J.M.; Kramer, S.; Stoll, D.; Templin, M.F.; Joos, T.O. Protein microarrays: Catching the proteome. Mech. Ageing Dev. 2005, 126, 161–170.

[11] Hall, D.A.; Ptacek, J.; Snyder, M. Protein microarray technology. Mech. Ageing Dev. 2007, 128, 161–167.

[12] Lopez, J.E.; Beare, P.A.; Heinzen, R.A.; Norimine, J.; Lahmers, K.K.; Palmer, G.H.; Brown, W.C. High-throughput identification of T-lymphocyte antigens from Anaplasma marginale expressed using in vitro transcription and translation. J. Immunol. Methods 2008, 332, 129–141.

[13] Anderson, K.S.; Ramachandran, N.; Wong, J.; Raphael, J.V.; Hainsworth, E.; Demirkan, G.; Cramer, D.; Aronzon, D.; Hodi, F.S.; Harris, L.; et al. Application of protein microarrays for multiplexed detection of antibodies to tumor antigens in breast cancer. J. Proteome Res. 2008, 7, 1490–1499.

[14] DeLuca, D.S.; Marina, O.; Ray, S.; Zhang, G.L.; Wu, C.J.; Brusic, V. Data processing and analysis for protein microarrays. Methods Mol. Biol. 2011, 723, 337–347.

[15] U. Alon, N. Barkai, D.A. Notterman, K.Gish, S. Ybarra, D. Mack, and A.J. Levine. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array", Proc. Natl. Acad. Sci. USA, Vol. 96(12):6745-6750, June 1999

[16] H. Zantema, S. Wagemans, D. Bo sˇnacˇ ki, "Finding Frequent Subgraphs in Biological Networks Via Maximal Item Sets", Bioinformatics Research and Development BIRD, Communications in Computer and Information Science, Springer, vol. 13, (2008), pp. 303-317.

[17] Chandra H, Reddy PJ, Srivastava S (2011). Protein microarrays and novel detection platforms. Expert Rev Proteomics, 8, 61-79.

## AUTHORS

Mr. N.Sevugapandi., MCA., M.Phil., (Ph.D).,is working as an Assistant Professor in Department of Computer Science at Sri Kaliswari College, Sivakasi. He has 6 years of teaching experience and has published more number of research articles in various National & International Journals. He is pursuing his Ph.D. at Bharathiyar University, Coimbatore in computer science discipline under the specialization of Data mining and Bio Informatics.

Ms.M.Hemalatha has 4 years of experience in teaching. She is an Assistant Professor at SubbaLakshmi Lakshmipathy College of Scince, Madurai. She graduated from Madura College, with bachelors degree in Mathematics in 2008. After that she completed post graduate in Master of Computer Application at Madurai Kamaraj University. She has keen interest in presenting papers in national and international conferences. She has organized many workshops on ‘Mobile Apps and Web Apps development’ for the Computer Science students.