Journal of Animal & Plant Sciences, 2010. Vol. 6, Issue 3: 645- 652.
Publication date: 28/04/2010, *http://www.biosciences.elewa.org/JAPS;* **ISSN 2071 - 7024**

An overview of advances in bioinformatics and its application in functional genomics

-Review paper-

Mwololo, J.K., Munyua, J.K., Muturi, P.W.and Munyiri, S.W.
[1]Makerere University, Faculty of Agriculture P.O. Box 7062 Kampala, Uganda; [2]University of Nairobi, Biochemistry Department, P.O. Box 30197, 00100 Nairobi, Kenya.

Corresponding author: mwololojames@yahoo.com

Key words
Bioinformatics, functional genomics, genes, microarray, biological, DNA,

1 SUMMARY

Bioinformatics is the scientific discipline that is concerned with the efficient management and useful interpretation of large scale biological information. Functional genomics aims at mapping DNA sequences and the components they encode for, to the function they perform. Initial efforts in bioinformatics were focused on the analysis of DNA sequence data. Presently, the scope and objectives of bioinformatics research and development have been broadened owing to the accelerating generation of data from various sources and for various cellular processes, the continuously evolving analytical technologies and the increasing computational capability. Bioinformatics offers an indispensable technology for function assignment and it has been used widely for gene annotation based on protein function predictions. However, as the sequence information is growing exponentially, the number of genes of unknown function is also growing, creating a challenge in the current computational approaches applied in bioinformatics. These limitations are being overcome through advances combining experimental and computational approaches, e.g. nanofabrication techniques. Despite the progress attained, analysis frameworks that could be used to analyze large data arising from signal transduction and biotransformation to provide quantitative predictions are inadequate. Trancriptome profiling is important because it provides information on the number of genes and their abundance in a tissue or given an induced condition e.g. diseased plants. Microarrays are hybridization experiments involving comparison of relative amounts of cellular mRNA from two tissue samples. Most of microarrays used in biological sciences can be divided into complementary DNA (cDNA) and oligonucleotide microarrays. The exploitation of hybridization in microarray analyses has sharply accelerated the search for defective genes of interest in both plants and animals. Microarrays provide the means to repeatedly measure the expression levels of a large number of genes at a time. Major limitations of this technology include decreased sensitivity of the arrays to the detection of genes with low expression levels and difficulties in data exchange due to the lack of standardization in platform fabrication, assay protocols and analysis methods.

Journal of Animal & Plant Sciences, 2010. Vol. 6, Issue 3: 645- 652.
Publication date: 28/04/2010, http://www.biosciences.elewa.org/JAPS; **ISSN 2071 - 7024**

2 INTRODUCTION

Bioinformatics is the development and implementation of computational tools and frameworks for the management, analysis and interpretation of biological information, whereas functional genomics is the integration of analytical technologies and bioinformatics for the elucidation of the function of cellular components (Vassily, 2000). Functional genomics has emerged as a research field that aims to map DNA sequences and the components they encode for, to the function they perform within various cellular processes. Modern technologies have allowed high-throughput generation of information about the DNA sequence of the genomes of an organism, the quantitative monitoring of its RNA and protein molecules, the identification of protein-protein and DNA-protein interactions and the mapping of the genetic variations within a population. Computer science, statistics, and biology have given birth to bioinformatics as a new discipline that is concerned with the efficient management and useful interpretation of large-scale biological information (Schaffer *et al*., 2000).

Early efforts in bioinformatics were focused on the analysis of DNA sequence data. This involved the design and integration of DNA sequence databases, the alignment of protein and DNA sequences, the assembly of DNA sequence fragments into genomic maps and the prediction of the function of a gene based on comparison of its sequence with sequences of genes with known function (Botchner *et al*., 2001). The scope and objectives of bioinformatics research and development have now been broadened due to the accelerating generation of data from various sources and for various cellular processes, the continuing improvements in analytical technologies and the increasing computational power of modern bioinformatics tools. This is because researchers are actively involved in solving new emerging issues such as plant and animal diseases, consequently leading to adoption of molecular techniques in expediting the process. Prediction of protein structure, image analysis, data visualization methods, analysis of gene and protein expression data, and simulation and dynamic analysis of integrated cellular processes are some of the increasingly important areas of bioinformatics (Adams, 2008). The identification of the elementary functions of genes has been a main field of study in biology and biochemistry research. The need for improved understanding of the systemic properties of genes has emerged due to constraints encountered in drug discovery, metabolic engineering, agricultural biotechnology and molecular biology. Identification of the intertwined functions of genes in metabolic processes in living organisms is essential and bioinformatics offers an indispensable technology for function assignment (Cebula, 2005).

The DNA sequence information is growing exponentially and each new DNA sequence is computationally analyzed to identify regions that might encode for proteins (genes) and the regions that might be responsible for gene regulation of transcription (Quellette & Baxevanis, 1998). The next step is prediction of function of the various genes. This is done using pair wise alignment algorithms that compare the DNA sequence of new genes with DNA sequences of known functions from other organisms (Pavlidis *et al*., 2002). The output of these computational analyses is a similarity index between every new gene and every gene in the databases. The extent of similarity between two sequences is based on the sequence identity and or conservation. Thus, if a new gene is similar to a gene of known function, the new gene might perform the same function. The genome-wide analysis that assigns a possible function to every new gene is called genome annotation, and

bioinformatics is being widely applied in this aspect (Jung and Thon, 2006).

3 CHALLENGES AND ADVANCEMENT

In every new genome some of the genes cannot be assigned a function with confidence using the current computational approaches. Moreover, a significant percentage of the unassigned genes do not share significant similarity with any other known gene (present in databases). This implies that as the sequence information is growing exponentially, the number of genes of unknown function is likely to increase substantially. A number of intelligent computational approaches have been developed to narrow this gap and one of them employs hidden Markov models for the identification and classification of DNA sequences within a gene that might be responsible for a specific protein domain (Durbin *et al*., 1998). In this approach, a hypothetical function can be assigned to a protein if it shares common domains with proteins of known function even if, based on pair wise comparison methods, it is not significantly similar to them.

There are also limitations in the ability of current computational approaches to predict the function of a gene and its product based on DNA sequence information. A combination of experimental and computational approaches is needed to overcome these limitations coupled with development of new more powerful computing technologies. In most cases, computational analysis of DNA sequence provides information about the elementary function of a gene and its product. Recent developments in analytical methods allow the observation of a set of genes and their products "in action" providing information about their systemic function i.e. transcription, translation, protein modifications and their functions there of. Using nanofabrication techniques, One of the technologies used to infer the function of various genes and to understand the complex interactions between cellular function and environment allows rapid and cell-wide monitoring of gene expression profiles, whereby thousands of DNA probes are attached to microarrays, and they hybridize with fluorescent labeled complementary DNA (cDNA) that has been quantitatively produced from the messenger RNA (mRNA) content of whole cells (Nature Genetics, 1999).

Review of the current technologies for cell-wide monitoring of the cellular processes suggests that as we move from DNA sequence through transcription and translation to signal transduction and biotransformation, the technologies are limited. On the other hand, analysis frameworks that could be used to analyze these data and provide quantitative predictions are also inadequate.

3.1 Use of transcriptome and proteomic tools to solve biological problems – the case of microarray technology in functional genomics: Molecular genetic studies, in combination with the extensive new body of sequence information for different genomes have revolutionized how cellular processes are investigated (Lockhart & Winzeler, 2000). New types of experiments are possible and discoveries are being made on an unprecedented scale. Trancriptome profiling is important because it provides information on the number of genes and their abundance in a tissue or given an induced condition, for instance in diseased plants in depicting mechanisms underlying gene-for-gene resistance and basal defence, host vs. non-host resistance among other environment induced conditions (Wise *et al*., 2007). A number of functional methods for determining gene expression have been developed, including microarrays, total gene expression analysis (TOGA), massive parallel serial sequencing (MPSS), subtraction hybridization (SBF) and serial analysis of gene expression (SAGE). This review focuses on micro arrays that involve synthetic oligonucleotides or complementary DNA sequences immobilized on membranes or solid surfaces.

Microarray is an approach of analysing thousands of genes. Microarrays are hybridization experiments involving comparison of relative amounts of cellular mRNA from two tissue samples. The terms "hybridize" and "hybridization" means that a single strand of DNA or RNA consisting of unpaired nucleotide bases bonds to a respective complementary nucleotide strand of DNA or RNA. Genomic DNA is usually first transcribed into mRNA in the cell nucleus and subsequently translated into proteins in the cell cytoplasm. Total RNA is extracted from the tissue, and the quality of the total RNA is verified by electrophoresis and spectrophotometry. The mRNA is labelled and hybridized to the array which is of known function for quantification. This is achieved by introducing a fluorescent marker during the preparation of mRNA that can be detected and quantified by a laser scanner.

Most of the microarrays used in biological sciences can be divided into two: complementary DNA (cDNA) and oligonucleotide microarrays (Schulze & Downward, 2001). The cDNA probes are usually products of the polymerase chain reaction (PCR) generated from cDNA libraries or genomic DNA, and are typically in excess of 150 nucleotides in length. On the other hand, synthetic oligonucleotides have a maximum length of around 80 nucleotides, thus conferring greater specificity among members of gene families (Lipshutz, *et al*., 2002). Array fabrication involves either spotting of pre-synthesized probes using highly precise robots, or *in situ* synthesis on glass slides. In both groups, mRNA is extracted, reverse transcribed to cDNA, labelled and hybridized to probes on the surface of the array (Lipshutz *et al*., 2002). Two fluorescent dyes allow cDNA from two treatment populations to be labeled with different colors. When mixed and hybridized to the same array, the differentially labeled cDNA results in competitive binding of the target to the probes on the array. After hybridization and washing, the slide is imaged using a scanner and fluorescence measurements are made separately for each dye at each spot on the array. This dual labeling enables the ratio of transcript levels for each gene on the array to be determined (Brown & Botstein, 1999) Specialized software and data management tools are then used for data extraction and analysis.

The exploitation of hybridization in microarray analyses has sharply accelerated the search for defective genes for example in diseased plants. The hybridization is based on the Watson-Crick Model of base pairing of nucleic acids. Each probe on a microarray is designed to hybridize with unknown target mRNA. When samples labelled with fluorescent compound (dye) are applied to microarrays, hybridization take place between each probe and the target mRNA. Each microarray probe recognizes cDNA sequences by base pairing. After a series of washes to eliminate unbound nucleotides and nonspecific bindings, only the target probe complexes remain bound. The intensity of the fluorescent signal for each probe reflects the abundance of the target RNA in the sample.

4 APPLICATIONS OF MICRIOARRAY TECHNOLOGY

The microarray technology has been used widely in bioinformatics and functional genomics. Some of the applications include comparative genomic hybridization for assessment of genomic rearrangements, and messenger RNA (mRNA) or gene expression profiling, where this technology is applied in monitoring expression levels of thousands of genes simultaneously which is relevant to many areas of biology and medicine such as studying treatments, diseases and developmental stages. For instance microarrays can be used to identify disease genes by comparing gene expressions in diseased and normal cells (Angidotan & Perry, 2007). Other application include single nucleotide polymorphism (SNP) detection arrays, where it is used in looking for SNPs in the genome of populations; chromatin

immunoprecipitation studies, where it is used in determination of protein binding site occupancy throughout the genome; eevolutionary studies, which have gained prominence with the use of species-specific arrays in parallel; and use of microarrays in diagnostics in the context of disease outbreaks for which rapid diagnostic tools are needed (Stoughton, 2005).

4.1 Strength of using glass in micoarray technology: The most common solid supports for microarrays are glass microscope slides and the advantages are that glass has low intrinsic fluorescence, is chemically inert, and the use of slides of standardized dimensions simplifies manufacturing and handling. Another advantage of glass as a substrate for microarray manufacture is that it is possible to deposit capture probes in spots of very small size, allowing a higher density of capture probes than on substrates such as nylon membranes, which are commonly used for low-density hybridization experiments. The increased miniaturization on glass supports makes it possible to represent many thousands of probes on a single array (Schena *et al.*, 1995). Microarrays provide the means to repeatedly measure the expression levels of a large number of genes at a time. Relatively small amounts of total RNA can be analyzed.

4.2 Limitations of Microarrays: Traditional methods that measure gene expression (e.g. Northern blotting, RNase protection assays) provide high resolution and can be used to validate or extend microarray data, though they are labour intensive. A couple of limitations are associated with the micro array technology. The planar surfaces such as glass have a much lower capacity to bind and or capture probes than porous substrates such as nylon membranes, and consequently exhibit lower signal strengths and a relatively narrow dynamic range of the genes of interest being studied. A major limitation is a decreased sensitivity of the arrays to the detection of genes with low expression levels (low-abundance genes). Another disadvantage is that microarrays do not measure post-translational modifications (e.g. phosphorylation) (Luo & Geschwind, 2001). In addition, it is possible to confound microarray results through a process of cross-hybridization in which specific components of the arrays will cross-hybridize because of sequence similarity of the probes. This is actually more of a problem with spotted DNA arrays since there is no attempt to spot onto the arrays only on the portions of genes that are different from their family members. This problem can be avoided by using oligos corresponding only to the regions that differ in sequence between closely related genes hence the researcher can solve this problem by selecting such oligonucleotide sequences..

More over tissue heterogeneity remains a persistent challenge for microarray studies, particularly where there are multiple densely packed cell types. This is also true for Northern blots, SAGE, and any other non-*in-situ* method. The measures of heterogeneous cell expression may decrease the sensitivity of microarrays by masking changes in gene expression (Torres-Munoz *et al.*, 2001). To overcome this problem, laser-capture micro-dissection techniques are used that can measure the expression of single cells, providing the capability of isolating homogenous samples from heterogeneous blocks of tissue (Torres-Munoz *et al.*, 2001). Microarrays and laser-capture micro-dissection, used in parallel, provide complementary information concerning cell-specific gene expression changes that are representative of larger blocks of tissue.

Microarray data is difficult to exchange due to the lack of standardization in platform fabrication, assay protocols, and analyses methods i.e. microarray fabrication methods such as inkjet and microjet deposition or spotting technologies and processes, in situ or on-chip photolithographic oligonucleotide synthesis processes, and electronic DNA probe addressing processes; high-density microarrays for high-throughput screening applications and lower-density microarrays for various diagnostic applications etc

(Michael, J., 2002; Kusnezow *et al*., 2007; Bi Q *et al*., 2007). In the present there is no consensus on the best assay protocol to adopt but the emerging technologies such as the 'protein microarray prepared with phage-displayed antibody clones' have higher throughput and lower cost compared to previous technologies and are convenient in protein profiling of experimental and clinical samples. However there is need to standardize this and use instrumentation and data software analysis tools from the same provider to eliminate these shortcomings. The analysis of DNA microarrays poses a large number of statistical problems, including the normalization of the data. Normalization is due to variations generated by the many data point analyses at the same time. The differences in spot intensities between replicate arrays arise from normal experimental variation such as differences in growth conditions, number of cells in the culture, RNA isolation and label incorporation efficiency, hybridization efficiency and image acquisition, signal measurement accuracy and sensitivity (Hatfield *et al*., 2003). A strategy to counteract this involves use of internal references that allow normalization based on total or ribosomal RNA, house keeping genes or a reference RNA (Hatfield *et al*., 2003). Other drawbacks are that this technology cannot be used to identify novel genes, detect alternative transcripts and it cannot recognize small RNAs.

Microarray methods based on glass slide formats are also not well suited to high throughput testing, requiring a large amount of manual manipulation to complete the technique effectively. This type of approach has two distinct disadvantages in a routine testing laboratory: (i) the staff time required pushes up the cost of individual tests, and (2) it is not well suited to automation, so the method in this format is not suited to testing large numbers of samples.

5 FUTURE PROSPECTS

The ultimate goal of functional genomics and bioinformatics is to integrate large-scale data sets of cellular processes towards a quantitative and ultimately predictive understanding of the function of individual cells and multicellular tissues. The relative intensity of hybridized elements allows rapid, reproducible and parallel quantification of the mRNA species within a cell. Data mining, clustering, and statistical analyses are used to classify the expression profiles and to identify sets of genes that share similar expression patterns. Comparison of the gene expression profiles between two experimental conditions over time can provide important information about groups of genes whose transcription is subject to the same regulatory rules and, therefore, they *might* serve similar cellular objectives. The power of these methodologies in studying gene expression has attracted considerable attention (Bassett *et al*., 1999).

In the near future, DNA microarrays will provide a method for simultaneously monitoring levels of nearly every gene transcript in genomes of different living organisms. Advances in spotted DNA arrays include the greater availability and quality of full-length cDNA clones for spotting on chips. Longer oligonucleotides are also starting to be used with standard spotting technology. Other new and anticipated applications of micro arrays involve the study of binding sites for transcription factors on a genome-wide level (Shoemaker & Linsley, 2002). New discoveries in combinatorial chemical processing promises to advance microarray technology. These include new digital light processors and simplified synthesis of nucleic acids.

Current methods based on glass slides are reliant on approaches developed specifically for gene expression profiling. Diagnostics have differing requirements and some formats are becoming available that target gene annotation specifically increasing speed and throughput. Publication of such protocols that are for specific detection of known pathogens is

expected to increase the use of microarrays. Efforts based on detection of gene functions at a higher taxanomic level would increase the utility of testing to look for unknowns.

One of the major challenges to be resolved is that of sensitivity. Currently most techniques either amplify all the nucleic acid from the sample or involve the use of biased amplification techniques such as polymerase chain reaction with the associated problems of multiplexing. Methods are needed that can give specific amplification of the target organism (signal) against the background of the host material (noise).

Obviously, the future will bring new technologies for detecting plant pathogens, largely because of the current efforts in genomics and molecular biosystematics and because of new platforms that have been developed primarily in the field of clinical medicine and or in the field of biological warfare.

6 CONCLUSION

The microarray technology represents an unprecedented analytical tool for transcript profiling. Implemented in the context of a well-designed experiment, microarrays provide high throughput, simultaneous analysis of mRNA for hundreds, if not thousands, of genes (Aharoni & Vorst, 2002). Such comprehensive analysis provides the opportunity to explore molecular mechanisms that underlie a variety of plant physiological processes and therein cDNA and oligonucleotide arrays can be used to gain a direct link to gene function (Schena *et al*., 1998). Thus, by correlating changes in gene expression with changes in physiology, it should be possible to derive insight into a broad range of biological processes and abnormalities. Microarrays have already been used to characterize genes involved in the regulation of circadian rhythms, plant defense mechanisms, oxidative stress responses, fruit ripening, phytochrome signalling, seed development and nitrate assimilation (Aharoni & Vorst, 2002). The microarray technology has been applied in the study of gene expression to serve as a central technological platform in solving biological problems by giving breeders the right information. This demonstrate how DNA/RNA microarrays have enabled the plant scientists to ask questions that were not possible a few years ago and illustrates how the field of plant science is being expanded by the availability of this emerging technology.

As a concluding remark, expression profiling technologies, in combination with other genomic tools, will have substantial impact on our understanding of plant-pathogen interactions and defence signalling pathways

.

7 REFERENCES

Adams, J. 2008: The proteome: discovering the structure and function of proteins. *Nature Education* 1(3).

Agindotan, B. and Perry, KL: 2007. Macroarray detection of plant RNA viruses using randomly primed and amplified complementary DNAs from infected plants. *Phytopathology* 97:119–27.

Aharoni, A. and Vorst, O: 2002. DNA microarrays for functional plant genomics. *Plant Molecular Biology* 48: 99–118.

Bi Q, Cen X, Wang W, Zhao X, Wang X, Shen T, and Zhu S, 2007: A protein microarray prepared with phage-displayed antibody clones. *Biosens Bioelectron* 15; 22(12):3278-82.

Bochner, B. R., Gadzinski, P., and Panomitros, E. 2001. Phenotype microarrays for high-throughput phenotypic testing and assay for gene function. *Genome Research* 11:1246–1255.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth,

R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J. and Corcoran, K: 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology* 18:630-634.

Brown, P.O and Botstein, D: 2000. Exploring the new world of the genome with DNA microarrays. *Nature Genetics* 21:33–37.

Cebula, T. A., Brown, E. W., Jackson, S. A., Mammel, M. K., Mukherjee, A. and LeClerc, J. E. 2005: Molecular applications for identifying microbial pathogens in the post-9/11 era. *Expert Rev. Mol. Diagn.* 5:431–445.

Jung, J. and Thon, M. R. 2006**:** Automatic annotation of protein functional class from sparse and imbalanced data sets. *LNCS*, 4316:65–77

Kusnezow W, Banzon V, Schroder C, Schaal R, Hoheisel JD, Ruffer S, Luft P, Duschl A, Syagailo YV, 2007: Antibody microarray-based profiling of complex specimens: systematic evaluation of labeling strategies. *Proteomics* 7(11):1786-99

Lipshutz, R.J., Fodor, S.P. and Gingeras, T.R: 2002. High density synthetic oligonucleotide arrays. *Nature Genetics* 21:20–24.

Lockhart, D.J, Winzeler, E.A: 2000. Genomics, gene expression and DNA arrays. *Nature* 405:827-836.

Luo, Z. and Geschwind, D.H: 2001. Microarray applications in neuroscience. *Neurobiology Disesaes* 8:183-193.

Michael, J. 2001: DNA MICROARRAY TECHNOLOGY: Devices, Systems, and Applications. *Annual Review of Biomedical Engineering* Vol. 4: 129-153

Pavlidis, P., Weston, J., Cai,j. and Noble, W.S. 2002: Learning gene functional classifications from multiple data types. *Journal of Computational Biology*, 9(2):401–411.

Pongrac, J., Middleton, F.A., Lewis, D.A., Levitt, P., Mirnics, K: 2002. Gene expression profiling with DNA microarrays: advancing our understanding of psychiatric disorders. *Neurochemical Research* 27:1049-1063.

Schaffer R, Landgraf J and Perez-Amador M. 2000: Monitoring genome-wide expression in plants. *Current Opinion Biotechnology* 11: 162–167.

Schena, M., Helle, r R.A and Theriault, T.P: 1998. Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnology* 16: 301–306.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O: 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–470.

Schulze, A. and Downward, J: 2001. Navigating gene expression using microarrays: a technology review. *Nature Cell Biology* 3: E190–E195.

Shoemaker, D.D and Linsley, P.S: 2002. Recent developments in DNA microarrays. *Current Opinion in Microbiology* 5:334-337.

Stoughton, R. B. 2005: Applications of DNA microarrays in biology. *Annual Review of Biochem*istry Vol 74:53–82.

Torres-Munoz, J., Stockton, P., Tacoronte, N., Roberts, B., Maronpot, R.R., Petito, C.K: 2001. Detection of HIV-1 gene sequences in hippocampal neurons isolated from postmortem AIDS brains by laser capture microdissection. *Journal of Neuropathology Experimental Neurol* ogy 60:885-892.

Vassily, H. 2000: Bioinformatics and Functional Genomics: Challenges and Opportunities.