eHealth Beyond the Horizon – Get IT There
S.K. Andersen et al. (Eds.)
IOS Press, 2008

Addressing the Biomedical Informatics Needs of a Microarray Laboratory in a Clinical Microbiology Context

Guillermo LOPEZ-CAMPOS, Victoria LOPEZ ALONSO[1], Fernando MARTIN-SANCHEZ
Medical Bioinformatics Department National Institute of Health "Carlos III", Madrid, Spain

Abstract. For an effective integration of microarray-based technologies in clinical settings a number of contributions from biomedical informatics technologies and techniques are needed to facilitate the improvement of the phases of experimental design, image analysis, data management, annotation, and analysis. In this communication we briefly present the state-of-the-art in the application of biomedical informatics to laboratories conducting microarray experiments and how our unit is coping with these requirements imposed by the routine clinical work of the National Centre of Microbiology, a reference laboratory for the Spanish Health System.

Keywords: Biomedical informatics, microarray, bioinformatics clinical microbiology.

Introduction

Microarray technology is being integrated into basic biomedical research and is becoming a fundamental molecular monitoring tool in clinical microbiology settings, especially for diagnostic applications due to the possibility of multiplexing the simultaneous detection of several pathogens in a single reaction [1, 2, 3]. Most of these clinical applications are based in the development of DNA microarrays for the detection of specific genes or gene regions in pathogens, but others are based in protein (ELISA microarrays) or other immobilised molecules. Clinical microbiologists can also apply microarray technology from other approaches such as analysis of gene expression to monitor microbial metabolism and screening of regulons to study microbial response to drugs, environmental changes, genome organization, and evolutionary studies [4].

Many software development projects have emerged in response to the needs of postgenomic research projects focused on laboratory information management [5], scripting and programming language (bioperl, biopython, biojava, bioruby…) [6], database integration [7] and data models and ontology development [8].

Along with this increasing diversity of laboratory techniques and uses of microarrays there is an increasing need of specific biomedical informatics tools in some areas of the field of microarray applications in clinical microbiology that are not covered by the

[1]Victoria López Alonso, Institute of Health "Carlos III", victorialopez@isciii.es

solutions designed and developed for gene expression studies, mainly in cancer research.

The aim of this work is to identify some of the biomedical informatics needs detected in a microarray laboratory focused on clinical microbiology (Figure 1) and how these user requirements could benefit from different synergistic developments of bioinformatics and medical informatics techniques and methods [9].

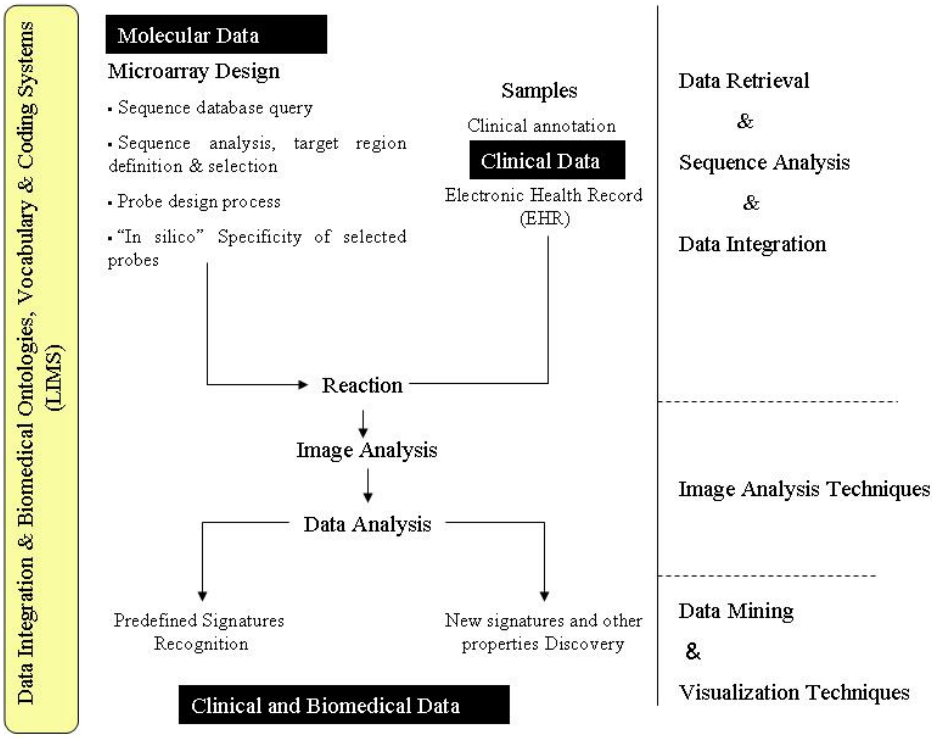

Figure 1. Biomedical informatics technologies of application in the different phases of a microarray experiment in a clinical microbiology context.

1. Material and Methods

1.1. Biomedical informatics and Microarray Design

Technological advances have boosted the field of microarray experiments. Although they represent high-throughput methods microarrays must be carefully designed in order to obtain useful data for clinical applications. Therefore design steps are key processes for the success or failure in the routine use of microarrays. In most of the clinical microbiological applications of microarrays, the design process is focused on the detection of probes that are capable to characterize and specifically identify the organism or organisms of interest. Therefore the initial step of design consists of

querying sequence databases and retrieving information. The target is the sequence of DNA to be amplified and fluorescently labelled. The strategy used to design it involves interrogating complete genome sequences stored in specific databases as the Comprehensive Microbial Genome [10] and expressed sequence tag databases available in the public domain. This step is problematic due in some cases to the vast amount and high redundancy of available sequence data or in other cases to the lack of available sequences in databases. At this point considerable effort has already been spent in generating EST libraries for the design of microbial genome microarrays.
Microarray probe design tools are very inefficient for large genomes and many existing tools operate in a batch mode. There are now several technologies that are useful for the design of microarrays for a selected set of model organisms based on Gene Ontology (GO) classification or using UniGene clusters [11].
Specificity of the designed oligonucleotides has to be checked with BLAST (Basic Local Alignment Search Tool) [12] and smart filtering techniques, which are employed to avoid redundant computation while maintaining accuracy.

1.2 Data management and annotation for Clinical Microarrays

Managing a microarray laboratory, especially for clinical use, typically involves good organization and sound tracking of experiments and samples data, including information on clinical samples, treatment, and experimental results. The use of Laboratory Information Management Systems (LIMS) addresses this problem by storing information with rigorous data entry mechanisms to prevent inaccuracies and to avoid redundancy. LIMS represent repositories of raw data associated with experiments, and are extremely useful as transient repositories before final submission of experimental microarray results to public databases, such as Gene Expression Omnibus (GEO) [13] or ArrayExpress [14].

1.3. Image Analysis Techniques in Microarray Experiments

In medical image processing the image content is often represented by features that are computed from the pixel matrix in order to extract features that support the development of reliable clinical diagnostic systems. In the case of microarray images there are some added difficulties for image analysis because many features are of abstract nature, as for instance those derived from a wavelength transform function.
After a reaction, microarrays are scanned and image analysis software determines the raw values using a data file that identifies the features and defines their dimensions and locations. Data stored in a tagged image file format (TIFF) is quantified by the image analysis software. To quantify microarray images it is necessary to ensure the quality of the signals and to prevent artefacts. Finally, the arithmetic mean or better median of the foreground and background pixels is calculated and data is obtained in the form of a table [15].

1.4. Microarray Data Mining Techniques

Before obtaining microarray results and applying sophisticated statistics, it is important to adequately normalize the data. A number of normalization methods have been proposed to correct systematic biases resulting basically from different amounts of

DNA used for labelling, different efficiencies of the Cy-3 and Cy-5 dyes in the labelling protocols, and different detection efficiencies of the dyes [16]. In gene expression experiments, fundamental patterns are extracted by several clustering methods like hierarchical clustering, self organizing maps and support vector machines. Lastly, the use of additional knowledge sources in the microarray data analysis process can improve the discovery and validation of clinical hypothesis. Information on sequence and structure, gene and protein interactions, function annotation and ontologies, or genetics and metabolic pathways can significantly complement any data analysis and improve its results.

2. Results

2.1. BUSSUB, a tool for designing clinical microarrays

The use of oligonucleotide microarray technology requires a very detailed attention to the design of specific probes to be spotted on the solid phase. They must have high sensitivity to detect a target gene and must be unique with high specificity. Parameters such as probe length, number of oligonucleotides, maximal distance from the 3 end, melting temperature range, threshold to reject secondary structures and prohibited sequences have to be considered in order to obtain an efficient design. In addition, the designed oligonucleotides must be computationally optimized to achieve greater specificity and uniformity and to perform optimally under the same melting temperature and other experimental conditions, reducing the noise due to cross-hybridization. Although there is an increasing number of publicly available genetic and molecular analysis programs for microarray oligonucleotide design, most of them use the same algorithm or criteria with only scarce variation [17]. None of them gathers the whole set of criteria. For this reason we have developed BUSSUB [3] an amplicon retrieval software that simplify and boost the process of recovering sequences contained between two given regions. BUSSUB is essential is the first stage of microarray design because this system is able to deal with high complexity files including several complete bacterial genomes.

2.2. AMANDA, Clinical Microarray Laboratory Information Management System

We have developed AMANDA [2] a LIMS specifically designed to store information about patient data, sample description, experimental resources, experimental parameters and conditions, and raw and processed hybridization results. The integration of this information into the microarray information system facilitates the processes of quality control and assessment of microarray experiments as well as further clinical interpretation of the results. AMANDA is compliant with the minimal information about a microarray experiment (MIAME) format [18]. This system is able to store, manage and visualize all the data and the information coming from several studies, to process it as a homogeneous dataset instead of multiple and separate sources

2.3 Image Analysis Software

Image analysis software used in our microarray laboratory carries out three fundamental tasks: gridding (to locate each spot on the slide), segmentation (to

differentiate the pixels within a spot-containing region into foreground) and information extraction. Information extraction includes two steps: spot intensity extraction and background intensity extraction. Spot intensity extraction refers to the calculation of fluorescent signal of the foreground from segmentation process, while background intensity extraction utilizes different algorithms to estimate the background signal due to the non-specific hybridization on the glass.
Developments in image acquisition, analysis and informatics technologies are ongoing and are expected to broaden the usefulness of DNA microarrays. For example Stanford MicroArray Database (SMD) [19] has the ability to store, retrieve, display and analyze the complete raw data produced by several additional microarray platforms and image analysis software.

2.4. Information of Clinical Microarray experiments

Most of the information resulting from diagnostic microarray experiments in a clinical microbiology context is qualitative (presence or absence of signal). In other cases as in gene expression or protein assays, data need to be analyzed in the context of clinical and epidemiological information in order to extract relevant knowledge useful for developing clinical solutions (Figure 2).

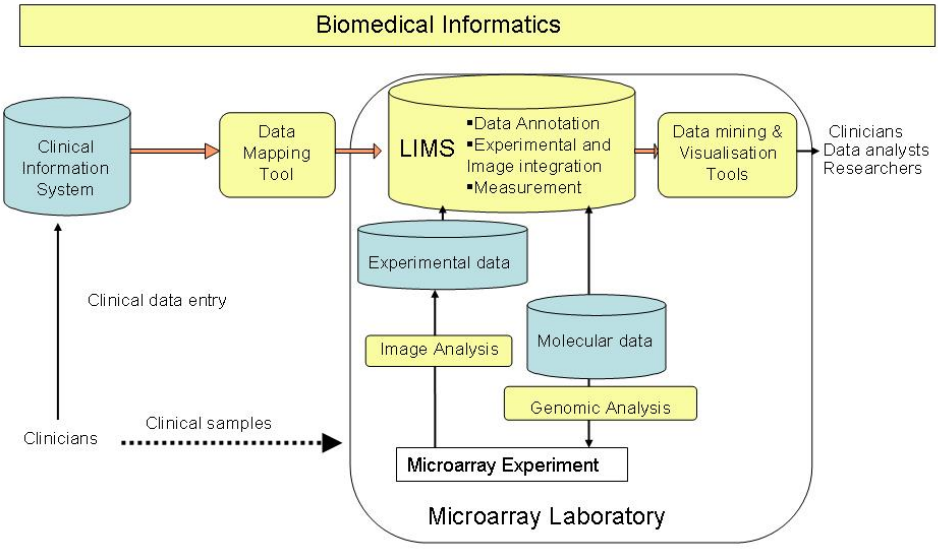

Figure 2. Implementation of biomedical informatics tools required for the production and use of clinical microarrays.

Data arising from microarray experiments can be normalized, filtered and analyzed utilizing several normalization and statistical modules available on integrative web based services. SMD [19] and the Bioconductor project [20] have added several tools and new ontologies to allow for an accurate and searchable annotation of biological

samples and experiments, developing and generalizing the schema for more efficient and flexible storage and analysis of microarray data.

Conclusions

High-throughput technologies such as microarrays are in the process of revolutionizing clinical microbiology. Further research and development of new biomedical informatics methods and techniques is becoming an integral part of a microarray laboratory. In our experience, these technologies are key in the process of assigning experimental genomic signatures to clinical profiles and to turn data into meaningful and reproducible clinical and mechanistic inferences.

References

[1] A. Loury and L. Bodrossy, "Highly parallel microbial diagnostics using oligonucleotide microarrays", *Clinica Chimical Acta*, vol. 363, pp. 106-119, 2006.
[2] G. López-Campos, L. Garcia-Albert, F. Martín Sanchez and A., Garcia-Saiz, "Analysis and management of HIV peptide microarray experiments", *Methods of Information in Medicine*, vol 45, pp. 158-162, 2006.
[3] G. López-Campos, M. Coiras, J.P. Sánchez-Merino, M.R. López-Huertas, I. Spiteri, F. Martín-Sanchez, P. Pérez-Breña, "Oligonucleotide microarray design for detection and serotyping of human respiratory adenoviruses by using a virtual amplicon retrieval software" *Journal of Virological Methods*, vol. 145, pp. 127-136, 2007.
[4] A. Ehrenreich, "DNA microarray technology for the microbiologist: an overview" *Applied Microbiology and Biotechnology* vol. 73, pp. 255-273, 2006.
[5] G. Thallinger, S. Trajaneski, G. Stocker and Z. Trajanoski, "Information management systems for pharmacogenomics", *Pharmacogenomics*, vol 3 pp. 651-667, 2002.
[6] http://www.open-bio.org
[7] V. Maojo, M. Garcia-Remesal, H. Billhardt, R. Alonso-Calvo, D. Perez-Rey and F. Martín-Sanchez,"Designing new methodologies for integrating biomedical information in clinical trials", *Methods of Information in Medicine*, vol 45, pp 180-185, 2006.
[8] http://www.obofoundry.org/ontologies.shtml
[9] D. Rebholz-Schuhman, G. Cameron, D. Clark, E. van Mulligen, J.L. Coatrieux, E. Del Hoyo Barbolla, F. Martín-Sanchez, L. Milanesi, I. Porro, F. Beltrame, I. Tollis and J. Van del Lei, "SYMBIOmatics: Synergies in Medical Informatics and Bioinformatics-exploring current scientific literature for emerging topics", *BMC Bioinformatics*, vol 8, (Suppl 1) S18, 2007.
[10] http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi
[11] "Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate", *Nucleic Acids Research*. vol 31, pp 3775-3781, 2003.
[12] http://www.ncbi.nlm.nih.gov/BLAST/
[13] T. Barret and R. Edgar, "Mining microarray data at NCBI´s Gene Expression Omnibus (GEO)", *Methods of Molecular Biology*, vol. 338, pp. 175-190, 2006.
[14] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans and A. Brazma, "ArrayExpress-a public database of microarray experiments and gene expression profiles", *Nucleic Acids Research*, vol. 35, pp. 747-750.
[15] A. Petrov, S. Shah, S. Draghici, S. Shams, " Microarray image processing and quality control" in S. Shah, G. Kamberova (eds) DNA array image analysis- nuts & bolts. DNA, Eagleville, pp 99-130
[16] J. Quackenbush, "Microarray data normalization ", *Nature Genetics*, vol 32, pp 496-501
[17] W. Rychlik "OLIGO7 " *Methods of Molecular Biology*, vol. 402, pp 35-60, 2007.
[18] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoecker, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton et al. "Minimum information about a microarray experiment (MIAME) - toward standards for microarray data", *Nature Genetics*, vol 29, pp 365-371, 2001.
[19] http://genome-www5.stanford.edu/resources/restech.shtml
[20] http://www.bioconductor.org