

# Instantiation and adaptation of CRISP-DM to Bioinformatics computational processes

Santiago González, Víctor Robles, José M. Peña and Ernestina Menasalvas

**Abstract**—Among the many contributions made by information technologies to Bioinformatics, the techniques of intelligent data analysis combined with optimization techniques are the main application field nowadays. Many researches focused on DNA microarray field have proposed different approaches trying to obtain new undiscovered knowledge of diseases such cancer. All these researches can be represented as a standard unique process. Thus, this paper presents an overview of a common biological and computational process of DNA microarray data analysis that include these types of researches, based on the known CRISP-DM model.

## I. INTRODUCTION

Nowadays, around 60 people die of diseases such as cancer every minute. The value is even more concerning if instead of thinking in minutes, we do it in hours or days. It is, therefore, a problem of high social impact that must be solved as quickly as possible. Finding a cure for diseases such as cancer would translate into a much higher life expectancy. In the scientific field, expert biologists are devoted to the study of possible solutions to these kinds of diseases. Among the many approaches, the DNA microarray technology will be the focus of this paper.

A DNA microarray is a large set of hybridized DNA molecules arranged on a solid (silicon or plastic) surface, called biochip. These types of experiments allow relative levels of mRNA abundance to be determined in a set of tissues or cell populations for thousand of genes simultaneously. A complete review of the methods used in the processing and analysis of gene expression for data generated by DNA microarrays experiments [11].

Many computer resources are needed in the work routine of an expert molecular biologist while studying DNA microarray data. That is why bioinformatics has been so important to meet the scientists' needs. The evolution of this new specialization was originally promoted by the biologists themselves and the needs they had at work. Nowadays, researchers from information technologies are beginning to work on this field, contributing on the data management and processing with their background of new tools and technology. We must bear in mind that we are talking of

rather complex information for non-biologists; therefore an intrinsic collaboration with the experts is absolutely essential.

Among the many contributions made by information technologies to bioinformatics, the techniques of intelligent data analysis combined with optimization techniques are the main application field nowadays. Many researchers contribute to improve, using these techniques, the results obtained with simple statistical studies. All researches that use Data Mining and Knowledge Discovery techniques to apply them on DNA microarray analysis are more or less supported on the same scheme or methodology. However, there is no any methodological process that describes all the possible Data Mining steps to analyze this kind of data.

Thus, this paper proposes an overview of a common biological and computational methodological process that includes practically all these types of researches. It is important to mention that the Computational Process is instanced and adapted of a known KDD model used by many Data Mining experts, which is called CRISP-DM.

The structure of the paper is as follows: The next section describes briefly the DNA Microarray technology. Section 3 presents the Biological process of DNA Microarray analysis, while section 4 describes the Computational process of these types of analysis and also analyzes each one of the steps of this process. Finally, all conclusions are presented in the last section (5).

## II. DNA MICROARRAY TECHNOLOGY

DNA microarrays [17,27,31,11] are a relatively new and complex technology used in molecular biology and medicine. Microarrays present unique opportunities in analyzing gene expression and regulation in an overall cellular context. This technology has been applied in diverse areas ranging from genetic and drug discovery to disciplines such as virology, microbiology, immunology, endocrinology and neurobiology. Microarray technology is the most widely used technology for the large-scale analysis of gene expression because it provides a simultaneous study of thousands of genes by single experiment.

A DNA microarray consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides (shorts molecules consisting of several linked nucleotides, between 10 and 60, chained together and attached by

Santiago, Victor and Jose M. belong to the Department of Computer Architecture, Universidad Politécnica de Madrid in Spain. (emails: {sgonzalez,vrobles,jmpena}@fi.upm.es). Ernestina is from the DLSIS Department, also at the Universidad Politécnica de Madrid, in Spain. (mail: emenasalvas@fi.upm.es)

covalent bonds), called Expressed Sequence Tags (ESTs), each containing several molecules of a specific DNA sequence. This can be a short section of a gene or other DNA element.

### III. BIOLOGICAL PROCESS OF DNA MICROARRAY ANALYSIS

There are several steps [28,21] in the design and implementation of a DNA microarray experiment (figure 1). Many strategies have been researched in each of these steps.

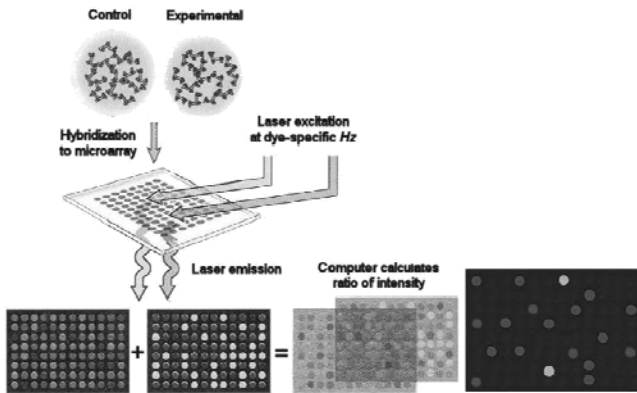


Fig. 1. Biological process of DNA Microarray analysis. Image from Gibson & Muse 2002

- **Probe:** First of all, the sample is obtained. The DNA type (cDNA/oligo with known identity) and the organism must be chosen in this step.
- **Chip manufacture:** The probes are placed on a surface. In standard microarrays, the information is attached to a solid surface by a covalent bond. The solid surface can be glass or silicon, in which case they are commonly known as gene chip or biochip. Here, several techniques have been used: Photolithography, pipette, drop-touch, piezoelectric (ink-jet), etc.
- **Sample preparation:** In this step the samples have been prepared. cDNA transcripts are prepared and labelled with a red fluorescent dye. A control library is constructed from an untreated source and labelled with a different fluorescent green dye.
- **Assay:** All information is hybridized (figure 2). Hybridization [28] is the process of combining single-stranded nucleic acids into a single molecule to the microarray.
- **Redaout:** Dual-channel laser excitation excites the corresponding dye, whose fluorescence is proportional to the degree of hybridization that has occurred. Relative gene expression is measured as the ratio of the two fluorescences: up-regulation of the experimental transcriptome relative to the control will

be visualized as a red pseudo-color, down-regulation show as green, and constitutive expression as a neutral black. The intensity of color is proportional to the expression differential.

- **Informatics:** In this final step, where new information and values are obtained from the fluorescence intensities using different computer techniques such as Robotics control, image processing [1], DBMS, etc. This step does not include data mining techniques, which have been studied as a computational process in next section.

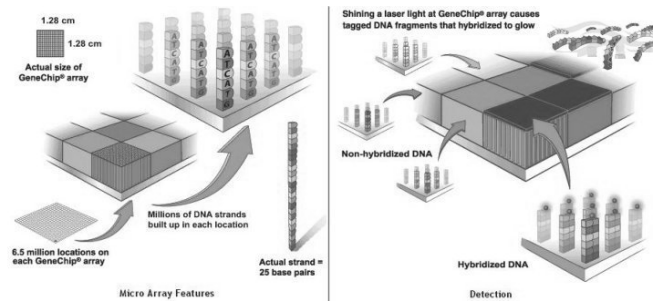


Fig. 2. Hybridization process. Image from <http://universe-review.ca/>

Nowadays, there are companies that create tools for analyzing complex genetic information such as DNA microarrays. Companies such as Affymetrix [7], Celera, Gene Logic, Xenometrix ... have built commercial platforms to carry out microarray experiments. Each platform obtains results using different methods (as Fluorescence, Mass spectrometry, Radioisotope, etc.) at each step of the microarray experiment. The use of platform determines the type of experimental design possible, the type of normalization, etc.

### IV. COMPUTATIONAL PROCESS OF DNA MICROARRAY ANALYSIS

Once Biological process is finished, the Computational process starts. Trying to obtain any standard methodological process that englobes all published researches, we propose to instance the CRISP-DM model [41]. This model distinguishes six main phases of a KDD process:

- **Business Understanding:** This initial phase focuses on understanding the objectives and requirements from a business perspective, then converting this knowledge into a problem definition and a preliminary plan designed to achieve the objectives.
- **Data Understanding:** The data understanding phase starts with an initial data collection and proceeds with activities in order to become familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets

to form hypotheses for hidden information.

- **Data Preparation:** The data preparation phase covers all activities for constructing the final dataset (data that will be fed into the modelling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed many times and not in any prescribed order. Tasks include record and feature selection as well as transformation and cleaning of data for modelling tools.
- **Modelling:** In this phase, various modelling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same problem type. Some techniques have specific requirements for the form of data. Therefore, stepping back to the data preparation phase is often necessary.
- **Evaluation:** Before proceeding to final deployment of the model, it is important thoroughly to evaluate the model and review the steps executed to construct the model in order to be certain it properly achieves the objectives. A key objective is to determine if there is some important issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the results should be reached.
- **Deployment:** Creation of the model is generally not the end of the project. Even if the purpose of the model is simply to increase knowledge of the data, the knowledge gained will need to be presented in a way that can be used.

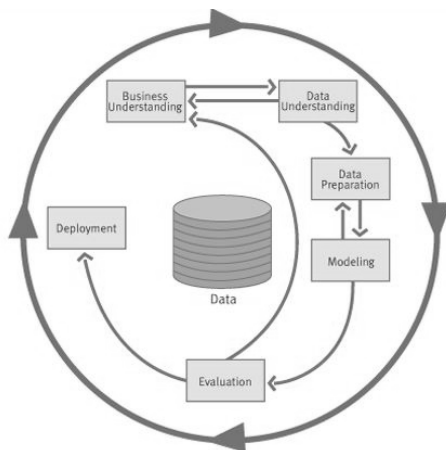


Fig.3. The phases of the CRISP-DM process model

Adapting this model to the microarray analysis process, the computational process of Microarray analysis is obtained. The Figure 4 shows a standard computational process with each phase. The life cycle of a computational process of DNA microarray analysis study consists of five phases. The sequence of the phases is not strict, moving back and forward between different phases is almost required, passing always on the Interpretation phase. It is because all decisions, results and objectives of each phase

have to be assessed and approved by expert biologists (biological interpretation). It is possible that an objective obtained in any phase has not a possible interpretation or is not a correct objective for the biologists. This can be produced, for instance, due to it is needed another Understanding iteration to understand the real objective. The lessons learned during the any phase can trigger new, often more focused questions to be answer by biologists.

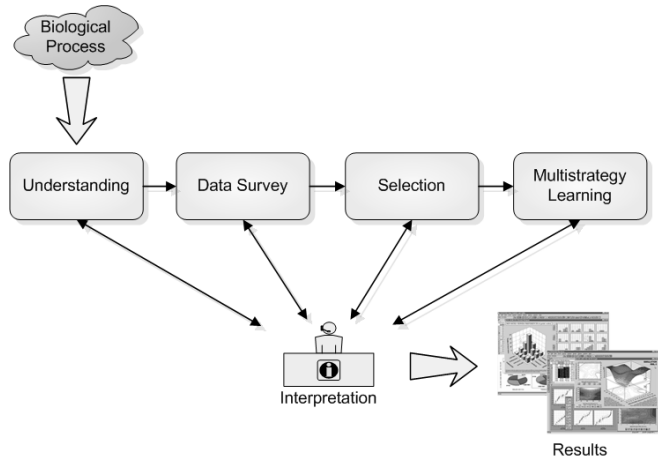


Fig. 4: Computational process of DNA Microarray analysis

A large number of data mining experiments with DNA microarray data can be represented with this methodological process, using all or not all phases, depending on the specific problem to be solved. In the next subsections each phase of the computational process is described and analyzed briefly.

### A. Understanding

This initial phase focuses on understanding the research objectives and requirements from the expert biologists, and then converting this knowledge into a data mining problem definition. The biologists define and comment one specific problem. They provide a microarray expression dataset with descriptions, headers, gene identifications, patient information and possible classification of diseases, labels or outcomes.

### B. Data Survey

In this phase all data is studied and prepared. Figure 3 defines the Data Survey tasks, Normalization and Pre-processing algorithms, both necessary to be able to access and compare correctly the data.

- **Normalization:** After the hybridizing and microarray image processing to obtain Cy5 and Cy3 fluorescence intensities (explained in section 3), it is needed to normalize [26,37,40] the data from each of the two scanned channels. There can be differences in labelling and detection efficiencies for the fluorescent labels and

differences in the quantity of the initial values from the two samples examined in the assay. These problems can cause a shift in the average ratio of the fluorescence intensities, so they must be re-scaled before an experiment can be properly analyzed. The normalization factor is used to adjust the data to compensate for experimental variability and to balance the fluorescence signals from the two samples.

There are many approaches for normalizing the gene expression. Some, such as total intensity normalization, are based on the assumption that the quantity of the initial RNA is the same for both labelled samples, so that consequently the total integrated intensity computed for all the elements in the array should be the same in both channels. Under this assumption, a normalization factor can be calculated and used to re-scale the intensity for each gene in the array. In addition to total intensity normalization, there are a number of alternative approaches for normalizing expressions, including linear regression analysis, log centering, rank invariant methods and Chen's ratio statistics (normalization using ratiostatistics), among others [26]. However, none of these approaches takes into account systematic biases that may appear in the data: dependence between intensity and ratio expression. Locally weighted linear regression (LOWESS) analysis [27], the most commonly used normalization method in DNA microarray experiments, can remove this dependency.

b) **Preprocess:** Obviously real data have a lot of redundancy, as well as incorrect or missing values, depending on some factors. Thus, usually it is needed some preprocessing algorithms in order to clean up and prepare the data. The most commonly used algorithms [17,8,38] are:

- Replicate handling or genes (features) that are replicated can be discarded.
- Missing value handling or patients (rows) that had more than 80% of missing gene values can be discarded.
- Imputing missing values can be estimated using different algorithms. The most known is the k-weighted nearest neighbor impute algorithm.

### C. Selection

In this phase a selection of the principal features is made in order to improve the understanding of the problem and its possible solution. In figure 3, Selection phase is divided in two possible tasks to obtain these features.

a) **Dimension Reduction:** Here a feature reduction task can be applied to the data. This task is used to discard features that are not relevant for the study or can produce noise. Among all the dimension reduction algorithms [4], the most broadly used ones in

microarray data [6] are based on Principal Components Analysis (PCA), Partial Least Squares (PLS) or even discarding variables with low internal variance or with low Pearson correlation with outcome. Other approach presents an algorithm based on Penalized Logistic Regression to make a dimension reduction [32].

b) **Feature Selection:** Trying to compare Feature Subset Selection (FSS) and Dimension Reduction, the first one selects only the best features from the data and the second one discards those features that are not relevant for the study.

FSS can be used as a simple task to obtain the best features to later obtain the best new knowledge using these features. For that, simple FSS algorithms based on statistics are proposed and compared in DNA microarray field [14], such as Fold Change, ANOVA, Rank Products, etc. However, FSS is a so important task in DNA microarray data that sometimes is the final objective of many researches, that is to obtain the Biomarkers. FSS in Bioinformatics are reviewed by Saeys [30]. These techniques use wrapper and filter mechanisms with supervised and non-supervised algorithms. Thus, although the final objective is Feature Selection, it is needed to execute a Multistrategy Learning phase which includes supervised and non-supervised learning.

### D. Multistrategy Learning

This phase is divided in two possible tasks (figure 3), unsupervised and supervised learning. Both tasks are used to obtain new knowledge (using different data mining algorithms) or the final objective of the process.

a) **Unsupervised Learning:** In DNA microarray technology, genes (features) classification is one of the typical final objective, although patients (rows) can be classified too. This classification can be obtained using different methods [18,25,34], such as Hierarchical cluster, EM, K-Means, QT, etc. Furthermore, it is interesting to obtain a patient classification (using the same methods) to later use this new information on the next task (supervised learning), to enrich knowledge and improve possible learners.

In Unsupervised Learning task, several studies obtain a Feature Subset Selection using wrapper mechanisms and unsupervised classification algorithm, such as EM or K-Means, to identify relationship between gene expressions [9,16]. Other researches [33] have proposed the use of biclustering technique (genes and patients simultaneously) to obtain better knowledge.

b) **Supervised Learning:** Usually in this task it is used a simple supervised learning using any supervised

classification method. Larrañaga [18] mentions the most used supervised classification methods in Bioinformatics, such as SVM, KNN, NaiveBayes, etc. However, this task can be something more complicated as a simple supervised classification algorithm. The same as Unsupervised Learning task, in Supervised Learning it is possible to make a FSS using wrapper [5] and filter mechanisms and different supervised classification algorithm, such as logistic regression [39,36], KNN, C4.5, NaiveBayes [12]. Furthermore, several researches use evolutionary algorithms, such as genetic algorithms [24], EDAs [29], or hybrid evolutionary algorithms [19] with supervised classification methods.

### E. Validation

Both Unsupervised and Supervised Learning tasks have to be externally validated. This validation is based on three aspects:

- **data mining** external validation, using a validation technique (depending on the classifier) and a external dataset. For supervised classification, cross-validation and bootstrap [3] have been the most commonly used validation methods, but [13] comments that these methods are unreliable in small sample classification. For unsupervised classification,
- **using literature** and comparing obtained results with other results.
- **using biological experiments** validations for validate our work or using gene databases, such as GO or KEGG.

### F. Interpretation

All interpretations, decisions, processes, feature selections and relations between genes or patients must be assessed and approved by expert biologists in order to obtain a valid result and/or objective in the microarray analysis.

## V. CONCLUSIONS

A complete standard biological and computational process of DNA microarray analysis is proposed. Mention that image processing techniques have been studied out of the computational process. Approaches, such as [35,15,20,23], etc., use each one of these phases, creating an overall computational process as the proposed. Other approaches fit with the proposed process in several phases. Applications, such as Bioconductor [10], GAMS [22], Knime [2], Weka, etc., allow us to create and execute each one of the steps proposed in this paper.

## ACKNOWLEDGEMENTS

The authors are grateful to the Blue Brain Project Team and the Canada IT-NRC Team, especially Fazel Famili, for their technical assistance. They also thankfully acknowledge the computer resources, technical expertise and assistance provided by the Centro de Supercomputación y Visualización de Madrid (CeSViMa) and the Spanish Supercomputing Network.

## REFERENCES

- [1] P. Bajcsy. An overview of dna microarray image requirements for automated processing. In CVPR '05, page 147, Washington, DC, USA, 2005. IEEE Computer Society.
- [2] M. R. Berthold, N. Cebon, F. Dill, T. R. Gabriel, T. Kötter, T. Meinel, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel. Knime: The konstanz information miner. In GfKL 2007. Springer, 2007.
- [3] U. M. Braga-Neto and E. R. Dougherty. Is cross-validation valid for small-sample microarray classification?
- [4] Miguel Carreira. A review of dimension reduction techniques. Technical Report CS-96-09, Dept. of Computer Science, University of Sheffield, January 1997.
- [5] K. Chrysostomou, S. Y. Chen, and X. Liu. Combining multiple classifiers for wrapper feature selection. IJDMMM, 1(1):91-102, 2008.
- [6] J. J. Dai, L. Lieu, and D. Rocke. Dimension reduction for classification with gene expression microarray data. Statistical applications in genetics and molecular biology, 5, 2006.
- [7] D. D. Dalma-Weiszhausz, J. Warrington, E. Y. Tanimoto, and C. G. Miyada. The affymetrix genechip platform: an overview. Methods in enzymology, 410:3-28, 2006.
- [8] S. Durinck. Pre-processing of microarray data and analysis of differential expression. Methods in molecular biology, 452:89-110, 2008.
- [9] Jennifer G. Dy and Carla E. Brodley. Feature selection for unsupervised learning. J. Mach. Learn. Res., 5:845-889, 2004.
- [10] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol, 5(10), 2004.
- [11] Wolfgang Huber, Anja Von Heydebreck, and Martin Vingron. Analysis of microarray gene expression data. In in Handbook of Statistical Genetics, 2nd edn. Wiley, 2003.
- [12] I. Inza, P. Larranaga, R. Blanco, and A. J. Cerrolaza. Filter versus wrapper gene selection approaches in dna microarray domains. Artif Intell Med, 31(2):91-103, June 2004.
- [13] A. Isaksson, M. Wallman, H. Göransson, and M. G. Gustafsson. Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recognition Letters, 29(14):1960-1965, October 2008.
- [14] Ian B. Jeffery, Desmond G. Higgins, and Aedin C. Culhane. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics, 7:359+, July 2006.
- [15] K. Kaufman and R. Michalski. Discovery planning: Multistrategy learning in data mining, 1998.
- [16] Yongseog Kim and W. Nick Street. Evolutionary model selection in unsupervised learning. Intelligent Data Analysis, 6, 2002.
- [17] S. Knudsen. A biologist's guide to Analysis of DNA microarray data. JohnWiley and Sons, 2002.
- [18] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles. Machine learning in bioinformatics. Brief Bioinform, 7(1):86-112, March 2006.

- [19] A. LaTorre, J.M. Peña, S. González, O. Cubo, and F. Famili. Breast cancer biomarker selection using multiple offspring sampling. *Current Trends and Future Directions in ECML/PKDD 07*, 2007.
- [20] Seok Won Lee, Scott Fischthal, and Janusz Wnek. A multistrategy learning approach to flexible knowledge organization and discovery. In *In Proceedings of AAAI-97*, pages 15–24. AAAI Press, 1997.
- [21] D. J. Lockhart and E. A. Winzeler. Genomics, gene expression and dna arrays. *Nature*, 405(6788):827–836, June 2000.
- [22] Dugas M., Weninger F., Merk S., Kohlmann A., and Haferlach T. A generic concept for large-scale microarray analysis dedicated to medical diagnostics. *Methods of information in medicine*, 45:146–152, 2006.
- [23] A. Naderi, A. E. Teschendorff, N. L. Barbosa-Morais, S. E. Pinder, A. R. Green, D. G. Powe, J. F. R. Robertson, S. Aparicio, I. O. Ellis, J. D. Brenton, and C. Caldas. A gene expression signature to predict survival in breast cancer across independent data sets. *Oncogene*,
- [24] C. H. Ooi and Patrick Tan. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, 19(1):37–44, 2003.
- [25] Tuan D. Pham, Christine Wells, and Denis I. Crane. Analysis of microarray gene expression data. *Current Bioinformatics*, 1:37–53, 2006.
- [26] J. Quackenbush. Microarray data normalization and transformation - nature genetics.
- [27] J. Quackenbush. Computational analysis of microarray data. *Nat Rev Genet*, 6(2):418–427, June 2001.
- [28] J. Quackenbush. Computational approaches to analysis of dna microarray data. *Methods Inf Med*, 45 Suppl 1:91–103, 2006.
- [29] V. Robles, C. Bielza, P. Larrañaga, S. González, and L. Ohno-Machado. Optimizing logistic regression coefficients for discrimination and calibration using estimation of distribution algorithms. *TOP*, 16(2):345–366, December 2008.
- [30] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, August 2007.
- [31] M. Schena, R. A. Heller, T.P. Theriault, K. Konrad, E. Lachenmeier, and R.W. Davis. Microarrays: biotechnology’s discovery platform for functional genomics. *Trends Biotechnol*, 7(16):301–306, July 1998.
- [32] Li Shen and Eng C. Tan. Dimension reduction-based penalized logistic regression for cancer classification using microarray data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2(2):166–175, April 2005.
- [33] Q. Sheng, Y. Moreau, and B. De Moor. Biclustering microarray data by gibbs sampling. *Bioinformatics*, 19:196–205, 2003.
- [34] Q. Sheng, Y. Moreau, F. De Smet, K. Marchal, and B. De Moor. Advances in cluster analysis of microarray data.
- [35] G. Sherlock. Analysis of large-scale gene expression data. *Brief Bioinform*, 2(4):350–362, December 2001.
- [36] S. K. Shevade and S. S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.
- [37] G. K. Smyth and T. Speed. Normalization of cdna microarray data. *Methods*, 31(4):265–273, December 2003.
- [38] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, June 2001.
- [39] G. Weber, S. A. Vinterbo, and L. Ohno-Machado. Multivariate selection of genetic markers in diagnostic classification. *Artificial Intelligence in Medicine*, 31(2):155–167, 2004.
- [40] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4), February 2002.
- [41] O. Marban, J. Segovia, E. Menasalvas, C. Fernandezbaizan, *Toward Information Systems*, Vol. 34, No. 1. (March 2009), pp. 87–107. doi:10.1016/j.is.2008.04.003.