

# Proceedings: The Applications of Bioinformatics in Cancer Detection Workshop

IZET M. KAPETANOVIC,<sup>a</sup> ASAD UMAR,<sup>b</sup> AND JAVED KHAN<sup>c</sup>

<sup>a</sup>*Chemopreventive Agent Development Research Group,* <sup>b</sup>*Gastrointestinal and Other Cancer Research Group, Division of Cancer Prevention,* and <sup>c</sup>*Oncogenomics Section, Pediatric Oncology Branch, National Cancer Institute, Bethesda, Maryland, USA*

**ABSTRACT:** The Division of Cancer Prevention of the National Cancer Institute sponsored and organized the Applications of Bioinformatics in Cancer Detection Workshop on August 6–7, 2002. The goal of the workshop was to evaluate the state of the science of bioinformatics and determine how it may be used to assist early cancer detection, risk identification, risk assessment, and risk reduction. This paper summarizes the proceedings of this conference and points out future directions for research.

**KEYWORDS:** bioinformatics; data mining; cancer; early detection; risk assessment; genomics; proteomics; drug discovery

## INTRODUCTION

The Division of Cancer Prevention (DCP) of the National Cancer Institute (NCI) sponsored and organized the Applications of Bioinformatics in Cancer Detection (ABCD) Workshop on August 6–7, 2002. Speakers included representatives from government, academia, and industry in the area of bioinformatics as applied or applicable to cancer prevention. The goal of the workshop was to evaluate the state of the science of bioinformatics and determine how it may be used to assist early cancer detection, risk identification, risk assessment, and risk reduction. In the context of the workshop, a broad definition of bioinformatics was employed, that is, application of computer processes to solve biological problems; or, as defined on the NCI Web site, “bioinformatics is the development and application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those required to acquire, store, organize, archive, analyze, or visualize such data” (<http://otir.cancer.gov/tech/bioinformatics.html/>).

Recent technological advances in biology and biomedical areas are resulting in a large accumulation of complex and multivariate data, and the problem is how to optimize and make most efficient use of this deluge of information. A systematic approach to data collection, storage, analysis, and representation is needed. Advances in theoretical and computational tools are providing opportunities for thorough data

Address for correspondence: Izet M. Kapetanovic, Chemopreventive Agent Development Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892-7322. Voice: 301-435-5011; fax: 301-402-0553.  
kapetani@mail.nih.gov

**Ann. N.Y. Acad. Sci. 1020: 1–9 (2004). © 2004 New York Academy of Sciences.  
doi: 10.1196/annals.1310.002**

mining in these vast collections of information. Data mining is a process of extracting “useful” information from a large collection of data. The goal is to enable and facilitate the process of going from data collection to knowledge acquisition. However, this path from data collection to knowledge acquisition is difficult and uncertain due to the intricacies of biological systems, the multivariate nature of data, and an overabundance of complex, nonlinear relationships. In the area of cancer prevention, we would like to be able to identify signs and risks early on and prior to onset of clinical disease and thereby be able to intervene appropriately and in a timely fashion.

There are two general approaches to data mining: theory-driven and data-driven. Theory-driven approaches are more established, but require assumptions about the underlying relationships and more extensive statistical, mathematical, and computer knowledge. The newer data-driven methods, on the other hand, do not require a priori knowledge about the relevant theory or all possible nonlinear relationships, and they do not make assumptions about statistical distributions. The latter are more suited to finding hidden features in data where none are visible by conventional statistical methods or by human decision alone. They have learning and adaptive capability and the ability to handle imprecise and multivariate and multidimensional information. For example, data-driven methods have an ability to achieve nonintuitive, complex nonlinear separation between patient classes and they can help identify a disease pattern even if all individual biomarkers are within an established reference range or, conversely, identify a nondisease pattern when all individual biomarkers are outside of the known reference range. In another words, the relationship of these biomarkers to each other may be more important than their individual absolute levels. These computational techniques are also referred to as machine learning or artificial intelligence. They hold a promise to be able to process and analyze huge amounts of noisy data coming simultaneously from many different inputs. Recognizing the ever-growing role of computational methodologies in medicine, the FDA has issued guidance for this purpose: “Guidance for the Content of Premarket Submissions for Software Contained in Medical Devices” (<http://www.fda.gov/cdrh/ode/57.html/>).

Examples of bioinformatics tools and techniques include principal component analysis, hierarchical clustering, artificial neural networks, fuzzy logic, neuro-fuzzy logic, genetic and evolutionary algorithms, and support vector machines. A recent book<sup>1</sup> describes and discusses various data mining techniques. Their ability to reduce dimensionality and generalize through nonlinear approximation and interpolation is the common thread among them. In that, they can generate outputs from previously unseen inputs (unsupervised learning) or from previously learned inputs (supervised learning). These computational methodologies are also classified as “soft computing” (<http://www.soft-computing.de/>)<sup>2</sup> because they utilize tolerance for imprecision and uncertainty to produce new information via approximation. In essence, these techniques are general approximators of any multivariate nonlinear function. It is increasingly recognized that bioinformatics tools hold promise in early detection, risk identification, risk assessment, and risk reduction, thereby facilitating effective approaches for the chemopreventive intervention. Specific areas amenable to bioinformatics include the following:

- Pattern clustering
- Classification
- Gene and protein array analysis

- Image and signal processing
- Decision support
- Database mining.

A brief overview of commonly used bioinformatics tools and their applications was included in the workshop booklet and is discussed in some detail in the following chapter. In addition, newer, more innovative bioinformatics methods are discussed. Furthermore, specific bioinformatics techniques are discussed in detail, and present examples of their use in cancer early detection, risk assessment, and prognosis are provided.

Several general themes relating to the use of bioinformatics in medical research and clinical application were also addressed during the workshop. Issues that warrant consideration and further attention were identified. It was rightfully acknowledged that bioinformatics holds a significant potential to enable, facilitate, and expedite progress in early detection and risk identification, assessment, and reduction of cancer.

### EXAMPLES AND PROOF OF PRINCIPLE

Dr. Michael Bittner (National Human Genome Research Institute, NIH, Bethesda, MD) opened the workshop with a basic observation about cellular memory and inertia that was later also echoed by Dr. Arul Chinnaiyan (University of Michigan, Ann Arbor, MI). The idea presented was that cells, in general, display great inertia and normal cells exhibit very stable gene expression. Different tissues show different expression of unique genes and a different pattern of expression of more common genes. Perturbations of normal tissues tend to result in relatively minor changes in gene expression that are reversible on return of tissue to its normal state. This leads to a premise that phenotype differences should be detectable based on concerted changes in expression of genes beyond a certain threshold level. Cells tend to follow a chosen path until dysregulated, which then may lead to cancer. The goal here would be to identify highly discriminatory genes between different phenotypes, such as normal and diseased, diseased responsive to treatment and diseased unresponsive to treatment, susceptible to disease and not susceptible to disease, etc. For example, Dr. Bittner presented data showing WNT5A gene expression to be discriminatory and an important marker in human melanoma progression.<sup>3</sup> It was correlated with greater motility and invasiveness and appeared to have a strong correlation with the survival phenotype. A number of different bioinformatics approaches and applications were described during the workshop for distinguishing different cancer-related phenotypes.

In order to identify genes that best discriminate between normal and disease (cancerous) phenotypes, Dr. Joseph Ibrahim (University of North Carolina, Chapel Hill, NC) applied **Bayesian** method utilizing **Markov chain Monte Carlo** techniques for DNA microarray analysis. This method has the advantage of being able to handle small sample size and allow incorporation of other already available information such as historical data or expert opinion. Another Bayesian approach, **Bayesian Decomposition**, was employed by Dr. Michael Ochs (Fox Chase Cancer Center, Philadelphia, PA) to identify participating pathways based on changes in gene expression.<sup>4</sup> This approach works backwards, using observed changes in mRNA to discover changes in signaling that cause them. The ability of Bayesian Decomposition

to assign genes to multiple coexpression groups (multiple coregulation) and encode biological information into the system makes it very suitable for this task and overcomes the limitations of other systems. This approach holds promise in helping to identify errors in signaling pathways that can act as cancer triggering mechanisms and to evaluate how these signaling pathway errors are affected by intervention.

Dr. Margaret Shipp (Dana Farber Cancer Institute, Boston, MA) and Dr. Chinnaiyan demonstrated use of combined genomic, proteomic, and clinical data in their respective studies of cancer-related phenotypes. With the aid of supervised learning (weighted voting with cross-validation testing and support vector machine algorithms), Dr. Shipp was able to identify signatures of outcome in large B cell lymphoma and rational targets of intervention.<sup>5</sup> This study was able to correlate PKC-beta mRNA and protein expression with enzymatic activity and led to the design of a clinical trial with a selective PKC-beta inhibitor to decrease proliferation and increase apoptosis. Dr. Chinnaiyan combined the use of high-density DNA microarrays to identify candidate genes, tissue microarrays to validate the gene expression at the protein level, and clinical information to enable making an association in prostatic cancer.<sup>6</sup> Unsupervised average linkage hierarchical clustering of genes into benign and malignant clusters was employed, with further clustering of the malignant cluster. Using the above general approach, hepsin (a transmembrane serine protease) was shown to be upregulated in prostatic cancer, with the highest hepsin expression found in the precursor lesion, prostatic intraepithelial neoplasia (PIN).

The development, optimization, and use of a robust **artificial neural network** classifier to handle four different diagnostic categories (lymphoma, Ewing's sarcoma, rhabdomyosarcoma, neuroblastoma) of small round blue-cell tumors were described by Dr. Javed Khan (National Cancer Institute, NIH, Bethesda, MD). The classifier was optimized for sensitivity and specificity and used to identify the most important and relevant genes in this classification.<sup>7</sup>

**Proteomic fingerprinting** provides information that is complementary to genomic fingerprinting or phenotyping. Proteins impart cellular functionality as they carry out most of the work of the cells and also represent the majority of drug targets. Based on an assumption that there are hidden diagnostic signatures in serum, Dr. Emanuel Petricoin III (CBER, FDA, Bethesda, MD) applied bioinformatics tools to demonstrate that serum proteomic patterns reflect tissue pathologic states, as in the case of ovarian<sup>8</sup> or prostatic cancer. There are plans to also extend this approach to other organs. **Pattern recognition** and classification required less than 1  $\mu$ L of raw unfractionated serum. This approach utilized a supervised genetic algorithm to iteratively seek combination of mass-to-charge ( $m/z$ ) values that can be used to classify samples and unsupervised **Kohonen SOMs** (self-organized maps) as a fitness test. The system was designed to learn and adapt with new data. The results of the study led to a PMA (premarket approval) application to FDA requesting clearance to market a class III medical device. In addition to identifying proteomic patterns, it is also useful to identify proteins themselves. Dr. Vineet Bafna (Celera, Rockville, MD) described **SCOPE**, a probabilistic model for scoring tandem mass spectra against a peptide database.<sup>9</sup> This approach can be employed to identify and characterize proteins differentially expressed in diseased vs. normal tissues and thereby discover diagnostic markers and targets for intervention.

Dr. Robert Murphy (Carnegie Mellon University, Pittsburgh, PA) proposed a different approach based on differences in subcellular distribution of some proteins

between normal and cancerous tissues: “location proteomics”.<sup>10</sup> Supervised learning was used to generate a “class” of patterns for subcellular structures of interest from images of localization of different proteins, extract subcellular localization features (SLF, numerical values describing the distribution of the protein within the cells) independent of cell position and rotation, and enable classification methods to learn to distinguish classes based on the features. Projected use of this approach is to monitor dynamic properties of proteins, relate them to changes with disease state and therapeutic intervention, and apply them to screening, detection, and intervention.

In addition to genomic and proteomic areas, bioinformatics also plays a critical role in imaging. Dr. Matthew Freedman (Georgetown University, Washington, D.C.) described **CAD** (computer-aided detection) and **CADX** (computer-aided diagnosis) approaches to improve small lung cancer detection and evaluation of response to anti-estrogen therapeutic intervention for mammary tumors. Dr. Carlos Andrés Peña-Reyes (Swiss Federal Institute of Technology, Lausanne, Switzerland) described the **COBRA system** (computer-assisted case interpretation) for modeling the human decision process, not human reasoning, in breast cancer risk assessment based on mammograms.<sup>11</sup> It uses **fuzzy Co-Co** (cooperative coevolutionary) methodology, that is, two evolutionary algorithms, one searching for labels and the other for rules. This divide-and-conquer approach is believed to provide a better search power at a lesser computational cost. It holds promise in improvements in sensitivity and specificity.

In order to streamline, optimize, and expedite drug discovery and development process, *in silico* methods based on integration of biology, chemistry, medicine, and information technology were proposed. AnVil received two Small Business Innovation Research (SBIR) awards from NIH to develop computational tools for discovery of cancer drugs. Dr. John McCarthy (AnVil, Inc., Burlington, MA) described his company’s approach combining data mining, high-dimensional analysis and visualization, statistics, and domain expertise to convey information, extract knowledge, and identify structures, patterns, anomalies, trends, and relationships. Using their proprietary technology, they generated diagnostic predictor and treatment outcome predictor for AML and ALL leukemia based on genomic profile and proposed a 76-gene chip for these applications in personalized medicine. They also employed a similar approach for a lung cancer DNA chip and chemical structure-based predictive toxicology. Dr. Ganesh Vaidyanathan (DuPont de Nemours & Co., Wilmington, DE) described **InfoEvolve™**, a set of empirical modeling tools for transitioning from data to knowledge and based on **information theory** and **genetic algorithm**. Its advantages are that it can build models with both low bias (low errors during training) and low variance (low errors during validation) without requiring a compromise between the two. It can be used to discover important inputs, build predictive models along the line of information-weighted **Bayesian modeling**, and identify strategies for discovering and designing compounds with desired biological activities. This modeling approach was shown useful in drug discovery in that it allowed sampling of fewer compounds in order to get a certain fraction of active ones.

It is recognized that cancer is a complex and multifactorial collection of diseases and that individual variables (biomarkers and indicators of cancer phenotype) are not adequately predictive or discriminatory. Dr. Judith Dayhoff (Complexity Research Solutions, Inc., Silver Spring, MD) described integrated use of **multivariate analysis**, **artificial neural networks**, and complementary statistical tools in aiding early

cancer detection and risk assessment.<sup>12</sup> This approach employs composite medical index or panel of biomarkers (clinical, genomic, proteomic, biochemical, imaging, etc.) rather than individual variables. It is patient-specific in that individual patient's information is entered into a neural network and index reflecting that patient's status or risk is obtained. Several medical examples, including oncology, were presented. She also pointed out that the amount of data is not as important as data being sufficiently represented along the boundaries of the discrimination problem. Adding small amount of random noise to provide more data points around the boundaries for training neural networks is often helpful.

### SOFTWARE TOOLS

A need for open source software tools in public domain was recognized, and Sandrine Dudoit (University of California at Berkeley, Berkeley, CA) and John Quackenbush (The Institute for Genomic Research [TIGR], Rockville, MD) described two sets of resources for genomic microarray analysis at Bioconductor (<http://www.bioconductor.org/>) and TIGR (<http://www.tigr.org/software/>), respectively.

The **Bioconductor project**<sup>13</sup> makes available an open source and open development software to assist biologists and statisticians in the area of bioinformatics. Its primary emphasis is on inference using cDNA microarrays. It consists of several modules that facilitate the analysis and comprehension of genomic data and allow efficient representation and manipulation of large and complex data sets of multiple types. Bioconductor packages include tools for preprocessing cDNA microarray data (similar package is also available for the Affymetrix platform): **marrayClasses** (classes and methods for cDNA microarray data), **marrayInput** (data input for cDNA microarrays), **marrayNorm** (location and scale normalization for cDNA microarray data), and **marrayPlots** (diagnostic plots for cDNA microarray data); as well as tools for differential gene expression: **multtest** (multiple hypothesis testing) and **ROC** (receiver operating characteristic approach). In addition to making software tools readily available, it also provides a platform for rapid design and deployment of quality software.

Another set of tools is available at TIGR Web site (<http://www.tigr.org/software/>). There, one can find a variety of standard operating procedures and software tools that are freely available to the scientific community, including **ResourceRer** (database for annotating and linking microarray resources within and across species), **MIDAS** (Microarray Data Analysis System for microarray data quality filtering and normalization that allows raw experimental data to be processed through various data normalizations, filters, and transformations via a user-designed analysis pipeline), **MADAM** (Microarray Data Manager, to load and retrieve microarray data to and from a database), **MultiExperiment Viewer** (MEV, Java application designed to allow the analysis of microarray data to identify patterns of gene expression and differentially expressed genes, and providing large number of different data mining tools), and **Array Viewer** (software tool designed to facilitate the presentation and analysis of microarray expression data, leading to the identification of genes that are differentially expressed).

Other more specific tools are also available, such as **dChip** (DNA chip analyzer; <http://www.dchip.org/>), which was described by Dr. Cheng Li (Harvard University,

Boston, MA). The dChip is a tool for normalization and estimation of expression levels in multiple oligonucleotide experiments based on a multiplicative model.<sup>14</sup> It employs a probe-sensitivity index to capture the response characteristic of a specific probe pair from multiple chips and calculates model-based indices and thereby detects outlier probe sets. It also provides hierarchical clustering and **principal components analysis (PCA)**.

## STANDARDIZATION

In order to facilitate management, sharing, and mining of huge amounts of complex microarray data being generated, establishing of standards is of paramount importance. **Microarray Gene Expression Data (MGED) Society** (<http://www.mged.org/>) is an international organization aiming to establish and implement standards for microarray data annotation and exchange, including facilitating the creation of related databases and public repositories and development of data analysis tools. It aims to make huge amounts of genomic and proteomic data broadly accessible. **MIAME** (Minimum Information about a Microarray Experiment) is a defined standard or a set of guidelines outlining the *minimum* information required to unambiguously interpret microarray data and allow access and subsequent independent verification (<http://www.mged.org/workgroups/miame/miame.html/>).<sup>15</sup>

Several prominent journals, including *The Nature* and *The Lancet*, have recently endorsed MIAME as a standard requirement for authors submitting microarray data for publication. In addition, *The Nature* will also require submission of microarray data to a public database (<http://www.ebi.ac.uk/arrayexpress/> and <http://ncbi.nlm.nih.gov/geo/>).

There are also standardization efforts for protocols in microarray data analysis, such as the one by **CAMDA** (Critical Assessment of Microarray Data Analysis; <http://www.camda.duke.edu/>).

## MICROARRAY CROSS-PLATFORM META-ANALYSIS

Presently, genomic studies are most prevalent and common. In spite of this, these studies are limited to small data sets (limited number of samples) and data sets on different platforms (oligo, cDNA, ink-jet, etc.). In addition, technologies involved in genomic sample processing and analysis are changing and evolving and there are continual improvements of experimental protocols for samples and microarrays. Therefore, there is a need for building models across platforms and with combined data sets. This would provide a more general approach and more reproducible and comprehensive models and decrease system-specific biases and idiosyncrasies. In addition, it would provide opportunity to validate models using data from other laboratories and larger data sets. Dr. Chinnaiyan stressed a need to cross-validate or interstudy-validate multiple data sets *in silico* and demonstrated use of meta-analysis of microarray data to identify dysregulation pathways.<sup>16</sup>  $\alpha$ -Methylacyl coenzyme A racemase (AMACR), involved in  $\beta$ -oxidation of fatty acids, was identified as a possible tissue biomarker for prostate cancer (sensitivity of about 97% and specificity of about 100% in diagnosing prostate cancer needle biopsies) following meta-analysis

of 4 independent gene expression data sets.<sup>6</sup> Dr. Pablo Tamayo (Whitehead Institute, MIT, Cambridge, MA) described use of a **Large Bayes Inference with Relative Features** approach. It involves rescaling data from individual data sets, merging and normalization, relative feature extraction, discretization and selection, creation of a database of labeled frequent item sets (combinations of frequently observed features' values or common occurrences in the data), and building a Bayes classifier using a product approximation. The advantages of a Bayes classifier are that it works with a small number of data points, with missing values and features, and with large data dimensionality, and combines supervised and unsupervised methods. Application and benefits of this approach were demonstrated across platforms (leukemia and lymphoma subclasses) and combined data sets (4-class adenocarcinoma data set).

### FUTURE DIRECTIONS

Question and discussion sessions were lively and stimulated a lot of thought. Comments and recommendations tended to be general in nature—not specific, but applicable to cancer prevention. The following “needs” were identified:

- Bioinformatics analysis should be incorporated into the experimental design from the beginning, not as an afterthought.
- More holistic approach should be employed incorporating genomic, proteomic, biochemical, pathological, and clinical data. Model development should be based on both clinical and molecular information. Different variables should be considered in relation to each other rather than independently.
- Standard operating procedures (SOPs) should be developed for collection, handling, storage, annotation (in computable form), and analysis of specimens. They should have portability and be independent of a specific laboratory.
- There is a need for more prospective studies.
- Larger well-annotated data sets should be established to get around the limitations associated with having the number of variables greater than the number of samples by several orders of magnitude. Annotation of data sets should be into functional categories that are relevant to cancer. Large data sets could be attained by combining data sets, using cooperative group studies with a large number of subjects, and using experimental animal studies to validate bioinformatics tools and as a proof of principle.
- Public repositories/warehouses should be established for samples (biorepositories), data, bioinformatics tools, and standardized data sets. These could then be used for validation and data mining.
- Further developments are needed in computational tools for data analysis across protocols, platforms, multiple data sets, and independent laboratories.
- Selection and implementation of “gold standard(s)” for validation are needed. RT-PCR was proposed as the most accepted standard in genomics, but its implementation has drawbacks. Another approach was to establish “reference laboratories” for validations.

- Publications should include adequate information to evaluate, reproduce, and *in silico* validate computational methods and results.
- Cross-training opportunities should be made available and encouraged, if not required, between biology/medicine and computation/bioinformatics.
- Funding should be made available for training, study sections for technology-driven (as opposed to hypothesis-driven) proposals, and SBIR projects.
- It would be worthwhile to establish and maintain a list of commercially and publicly available computational tools along with a brief description and general rating in terms of overall quality, capabilities, and usefulness.
- Greater effort is needed in development and validation of text mining algorithms.

#### REFERENCES

1. HASTIE, T., R. TIBSHIRANI & J. FRIEDMAN. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag. New York/Berlin.
2. KECMAN, V. 2001. *Learning and Soft Computing*. MIT Press. Cambridge, MA.
3. WEERARATNA, A.T., Y. JIANG, G. HOSTETTER *et al.* 2002. Wnt5a signaling directly affects cell motility and invasion of metastatic melanoma. *Cancer Cell* **1**: 279–288.
4. MOLOSHOK, T.D., R.R. KLEVECZ, J.D. GRANT *et al.* 2002. Application of Bayesian decomposition for analyzing microarray data. *Bioinformatics* **18**: 566–575.
5. SHIPP, M.A., K.N. ROSS, P.B. TAMAYO *et al.* 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8**: 68–74.
6. RUBIN, M.A., M. ZHOU, S.M. DHANASEKARAN *et al.* 2002.  $\alpha$ -Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. *JAMA* **287**: 1662–1670.
7. KHAN, J., J.S. WEI, M. RINGNER *et al.* 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **7**: 673–679.
8. PETRICOIN, E.F., A.M. ARDEKANI, B.A. HITT *et al.* 2002. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**: 572–577.
9. BAFNA, V. & N. EDWARDS. 2001. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* **17**(suppl. 1): S13–S21.
10. BOLAND, M.V. & R.F. MURPHY. 2001. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics* **17**: 1213–1223.
11. PEÑA-REYES, C.A. & M. SIPPER. 2000. Evolutionary computation in medicine: an overview. *Artif. Intell. Med.* **19**: 1–23.
12. DAYHOFF, J.E. & J.M. DELEO. 2001. Artificial neural networks: opening the black box. *Cancer* **91**: 1615–1635.
13. DUDOIT, S. & Y.H. YANG. 2003. Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. *In The Analysis of Gene Expression Data: Methods and Software*. Springer Pub. New York.
14. LI, C. & W.H. WONG. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**: 31–36.
15. BRAZMA, A., P. HINGAMP, J. QUACKENBUSH *et al.* 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* **29**: 365–371.
16. RHODES, D.R., T.R. BARRETTE, M.A. RUBIN *et al.* 2002. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.* **62**: 4427–4433.