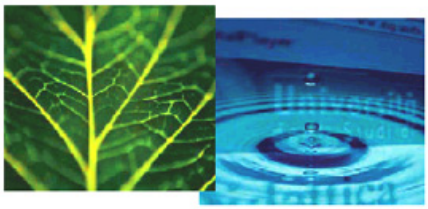
PhD Dissertation



International Doctorate School in Information and Communication Technologies

DIT - University of Trento

ON INFORMATION ORGANIZATION AND INFORMATION EXTRACTION FOR THE STUDY OF GENE EXPRESSIONS BY TISSUE MICROARRAY TECHNIQUE

Francesca Demichelis

Advisor:

Dott. Ing. Paolo Traverso

ITC-irst, Center for Scientific and Technological Research, Trento

February 2005

DIT-05-024

Abstract

Genomic expression studies are the means of depicting molecular profiles characterizing specific disease states. Microarrays allow the tracking and the translation of genome sequences into gene functions, leading to the identification of highly informative genes and pathways with a potential impact on understanding disease development and progression. These technologies concurrently may improve diagnostic and treatment modalities and the detection of novel therapeutic targets.

Expression array technology is dramatically expanding the amount of data available on many disease states. These studies typically involve many researchers with different backgrounds, each contributing to some steps of the entire process.

In particular, Tissue Microarray technology allows for high-throughput expression profiling of tumor samples by evaluating potentially interesting candidate genes and proteins on a large number of well-characterized tumors, providing information on a population basis.

High quality experimental data production is extremely important for the reliability of data analysis. Critical assessment of experimental design and organization and reliability assessment of experimental data together with data preprocessing need to be addressed. A technological approach is also advisable to properly manage data heterogeneity, data quantity and user diversity.

The focus of this thesis is to develop a systematic approach to processing and better understanding data generated from Tissue Microarray technology, overcoming the limitations of other current approaches. This thesis addresses Tissue Microarray data collection and organization, enhancing data sharing, usability, and process automation. We faced pre-processing issues, identifying critical points and some solutions. We also focused on a specific issue in data classification, proposing a novel classification model based on a Bayesian hierarchical approach, able to handle data uncertainty. Three Tissue Microarray experiments are pre-

sented as case studies with the purpose of providing real world examples to illustrate some of the critical points made in this thesis.

Keywords
Tissue Microarray, Expression micorarray, Molecular Profiling, Data Management

Contents

Chapter 1

1. Introduction

This doctoral thesis addresses how to handling gene and protein expression information, obtained by using a high throughput technique, called Tissue Microarray Technology. It encompasses the interdisciplinary field of bioinformatics. This chapter first introduces the headlines of this discipline.

Bioinformatics uses techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics and has many practical applications in different areas of biology, medicine and informatics. Generally speaking, all knowledge which can be extracted from computer analysis of biological data may be labeled as bioinformatics.
Therefore, as bioinformatics is really an heterogeneous and rapidly evolving field, it might be helpful to sub-divide its current possible different activities by key areas to organize concepts [1][2], keeping in mind that the individual research activity often embraces more than one area in a comprehensive transversal way. Bioinformatics could be divided into the following key areas:

I. Sequencing (Algorithms for DNA, genomic and protein sequences, Algorithms for sequence comparison and multiple sequence alignment)
II. Organization of Biological Knowledge (Development of database for bio-structure and bio-processes organization, Infrastructure for distributed resources, User Interface and Information Representation, Query Languages, Networks, Taxonomy and Ontology).
III. Knowledge Discovery & Data Mining (Algorithms for feature selection for gene reduction,

Learning algorithms for gene and protein expression).
IV. Biomolecular structure and Biological processes (Computational Structure Biology, Computational Systems Biology).

As suggested by Luscombe et al. [3] the development of bioinformatics techniques has essentially determined an expansion of biological research in two dimensions, each one with specific aims and related view of biological problems. Working in each of these two dimensions often requires the use of tools and knowledge specific to the four key areas of bioinformatics as defined above. The aim of the first dimension is to fully understand the information contained by individual genes. In this sense bioinformatics follows more the traditional reductionistic approach that was typical of biological studies. Starting with a gene sequence, the encoded protein sequence can be determined with strong certainty. Next step is to calculate the structure adopted by the protein. Finally, docking algorithms could design molecules that could bind the model structure, leading the way for biochemical assays to test their biological activity on the actual protein. So this dimension may be labeled as “rational drug process design”. In this context, *clinical validation* of a new gene potentially important in disease development and progression, as indicated in gene expression experiments, has a crucial role, allowing for the identification of novel therapeutic targets.
The second dimension refers information related to the systemic functional behavior of the cell or the organism. It is possible to distinguish an experimental and a theoretical approach. The former organizes biological knowledge principally by experimentally and computationally searching for similarities between different molecules. This allows for the transfer of information between related entities and eventually leads to gene discovery or protein families, which share structures and/or functions. In this way, biologists are able to compile a “genome census” that provides comprehensive statistical accounts of protein features, such

as the abundance of particular structures or functions in different genomes. Using this data it is also possible to trace the evolutionary path of proteins, gaining first insights into the evolutionary path of the whole organism. Expression microarray data represents a rapidly growing and exciting new source of genomic information. Grouping genes with similar expression profiles is a typical operation of this second dimension. In this way it is possible to help physicians improve disease classification leading to advances in diagnosis, prognosis and therapeutic treatment.

In theory, biological systems may be studied following a systemic approach, dividing biological information into two areas: (i) genes, proteins and individual molecules as basic components of biological systems; (ii) regulatory networks (which specify the expression patterns of genes and proteins), intracellular metabolic networks and both intra- and inter-cellular communication networks, in which molecules participate. This biological information is hierarchical. An important emerging challenge for biology and medicine is the study of complex biological systems by capturing and integrating these different levels of biological information, thus scaling up from molecular biology to systems biology. An integrated and more global view of the systems requires the application of models to hierarchically understand the interplay between atoms, molecules, regulatory and metabolic processes, and cells to help simulate or analyze dynamic behavior of biological systems. The ultimate goal of systems biology is to correctly understand the complex biological networks that define a cell response to environmental and genetic changes, therefore to correctly predict and modify the behavior of biological systems.

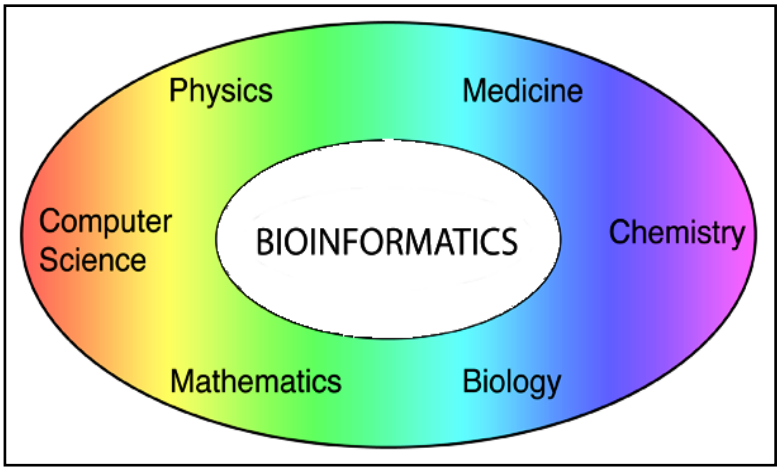


Figure 1 - Bioinformatics as interdisciplinary framework

1.1. The Context

Gene expression is the process by which genetic information is converted into the structures and functions of a cell, by producing proteins from DNA coding sequences. The amount of protein that a cell expresses depends on the tissue, on the developmental stage of the organism and on the metabolic or physiologic state of the cell. Molecular profiles, which represent the molecular fingerprint of cells, are investigated to depict profiles characterizing specific diseases [4].Molecular profiling studies are primarily conducted in a comparative manner. Molecular profiling can be based on comparison between non-diseased (i.e., normal) and diseased (e.g., tumors) samples, between diseased samples pharmacologically treated and untreated at variable time points or between samples of different diseases.

Expression array technology is dramatically expanding the amount of data available on many disease states. This technology allows tracking the translation of genome sequences into gene

functions (functional genomics). In the study of cancer, this new biological knowledge promises to improve diagnostic [5][6] and prognostic modalities [7][8][9] and to detect targets for new therapies [10].

Microarrays generate large molecular datasets in the setting of similar experimental conditions: DNA microarrays (DMAs) (first proposed in 1996, [11]) analyze mRNA expression, and Tissue microarrays (TMAs) (first proposed in 1998, [12]) evaluate DNA, RNA or protein targets through *in situ* investigations (analyses performed on tissues) by hybridization or immunhistochemistry. In molecular profiling studies of human cancer, integration of DMA and TMA experiments, in serial or cyclic processes, provides a powerful approach [13][14]. These studies typically involve many researchers with different backgrounds, each contributing to some steps of the entire process: biologists, chemists, pathologists, and clinicians. Their work is supported and complementarily completed by physicians, computer scientists, mathematicians, etc., who are required not just to perform data analysis, but also to contribute to the management of whole experimental process from design to final data interpretation.

As the throughput of this technology increases in the area of expression arrays, proteomics, and tissue microarrays, one of the critical issues is making sense of all this data. Therefore, the focus of this thesis is to develop a systematic approach to processing and bettering understanding data generated from one of these high-throughput discovery tools, the Tissue Microarray.
TMAs can be used to examine multiple disease states, but this thesis will focus on its uses in the area of cancer, where most of the initial work has been.

1.2. The Problem

Tissue microarrays allow for high-throughput expression profiling of tumor samples by evaluating potentially interesting candi-

date genes and proteins, identified by other techniques, on a large number of well-characterized tumors.
They can be used to investigate panels of molecules, trying to characterize solid tumor development or tumor progression. Moreover TMA studies have the potential to be translatable to a clinical application such as the development of diagnostic biomarkers (e.g., AMACR [15]) or a potential to therapeutic target (e.g., Her-2-neu [16]).

The TMA approach can be technically described as follows. It gathers into one paraffin block up to hundreds of minute tissue samples. Usually more than one sample from each patient is included (sample replicate) to ensure tumor representativeness. Slices of this 'block array' are then analyzed by in situ methods on glass slides, evaluating DNA, RNA or protein targets, for identification of specific phenotypic (immunohistochemistry and in situ hybridization) or genotypic (fluorescence in situ hybridization) alterations. Immunohistochemistry technique is most commonly performed, allowing the investigation of protein expression by analyzing the reaction of protein specific antibodies (biomarkers).

The TMA technique is thus appropriate for population screening studies. The information you can get are best viewed on a population basis and not on an individual patient basis. Thus the question being asked is not if patient A, B, or C demonstrates expression of the biomarker in question, but instead, how often is this biomarker expressed in a population of individuals with a given disease state? Is this a commonly expressed gene? Is it expressed only rarely? Is it expressed only in patients who go on to have a worse outcome?

In TMA experiments, as with all high throughput techniques, high quality experimental data production is extremely important for the reliability of data analysis. This critical issue needs to be addressed on two levels, i. critical assessment of experimental de-

sign and organization and ii. reliability assessment of experimental data together with data preprocessing. In order to best deal with these two issues, a technological approach is advisable to properly manage data heterogeneity (biological, clinical and technical variables), data quantity and user diversity. This last aspect becomes even more crucial if data sharing occurs among and between research groups. Technological aspects may include automation to speed up data acquisition and evaluation that can be of great impact in overcoming TMA studies bottlenecks.
Adequately tackling these problematic aspects is important to achieve good experimental results.
A fundamental phase of each study is then data analysis. Standard statistical approaches and supervised and unsupervised learning algorithms have been successfully applicable to TMA preprocessed data in order to face prognostic predictive power assessments, classification tasks or new class discovery tasks. However efforts are continuously done to better address peculiar necessities of these new types of data, by elaborating new dedicated methods for data analysis. For instance, some authors recently proposed approaches [17][18] to properly manage classification tasks when the target are continues data. This approach fits scenarios where the biological question is how can I characterize patients with good or poor prognosis equally histologically diagnosed/classified? Which is the profile that characterizes different clinical outcome data? Another issue that might be investigated regards the variability of measures in a classification problem. Standard approaches do not account for data uncertainty, which may be given by experimental multiple measures of protein levels on the same tumor (intra-tumor heterogeneity). Embedding this variability in classification models might add interesting information to data analysis outputs.

1.3. The Solution

To appropriately exploit the high-throughput nature of the TMA technique, a comprehensive approach to data management must

be applied. Knowledge Discovery and Data Mining (KDDM) [19] processes provide an ideal framework for bioinformaticians to fit: from understanding the domain, forming a consolidated dataset from data sources, cleaning the data by selection and preprocessing, extracting regularities in the data and formulating knowledge in the form of patterns and rules.

The solution we propose here is a systematic approach to processing and better understanding data generated in TMA based studies.

Through a strong interdisciplinary approach, we tried to address all the issues related to this topic. In particular, we aimed to support both data collection from different users and data sharing across different institutions through a web based approach [20]. To support automation we employed digital pathology for acquiring digital images of single tumor sample on a glass slide and assigning them to the donor tumor automatically [21] through image processing and object recognition algorithms. Fully automatic evaluation of specific biomarker categories have been developed in order to avoid subjectivity problem and fully exploit the variability range of gene expression [22].However while full automation may be the Holy Grail in these experiments, currently careful supervision is still required to ensure that the samples are properly classified.

This thesis addresses TMA data pre-processing issues, identifying critical points and some solutions. These models were tested on three TMA studies driven by different biological/clinical questions. The results of these three studies are presented in this thesis.

We also focused on the specific issue of accounting for data uncertainty in classification. We propose a novel classification model based on a Bayesian hierarchical approach. This model was tested on simulated data and on a real protein data set finding interesting differences, when compared to a standard Bayesian approach.

1.4. Innovative Aspects

We present a new comprehensive system for the management of TMA experiment data (TMABoost) from tissue data collection to experiment design and gene expression evaluation. This system is able to handle a wide spectrum of studies and user requirements. This system integrates all the information related to TMA experiments, regardless of experiment design, overcoming the limitations of the systems that have been presented up to now.
Our approach both reduces the possibility of errors and accelerates the data analysis step on reliable data, particularly with regards to the pre-processing phase. The system includes an image processing procedure and a robust algorithm for object recognition to automatically identify each tumor sample on the digital image of a TMA glass slide, assigning it to proper grid location, speeding up the acquisition of corresponding digital image by almost avoiding manual intervention. TMABoost allows data sharing among institutions and is also able to face data exchangeability needs [23]. An important aspect of our system is that, different from other academic systems or commercially proposed systems, it is patient based and not experiment based.
Another innovative aspect we propose regards the treatment of uncertainty in the context of classification. The classification model is based on Bayesian hierarchical approach. It allows embedding in the classification model the tumor intra-variability (heterogeneity of protein levels across tumor tissue), using the tuple of protein level measurements of each case instead of unique representative value, as done by conventional approaches.

1.5. Structure of the Thesis

The second chapter of this thesis reviews the state of the art of TMA field with respect to data management and data analysis tasks. The third chapter introduces problems of performing molecular profiling at protein expression levels, from data handling point of view. Insights in TMA experimental technique and *in situ*

investigation methods are provided to illustrate TMA data sources. The chapters 4, 5, and 6 describe the TMABoost management system, the TMA data analysis steps with particular attention to the preprocessing aspect, and the classification model we propose together with the performances we obtained. Chapter 7 includes three examples of TMA based studies we conducted, focused on different biological questions. The chapter 8 summarizes and discusses the problem and the proposed solutions and outlines on-going and future work on this topic.

1.6. Notes

Part of the work presented here has been granted by the Italian Ministry of Health ("Programmi speciali" – Art. 12 bis, comma 6, d.lgs. 229/99).
During the three years training period I applied for a three months stage at the CHIP (Children's Hospital Informatics Program, Boston, MA, http://www.chip.org/). During that period of time I collaborated with members of the Rubin laboratory (Mark A. Rubin, http://rubinlab.tch.harvard.edu/) at the Brigham and Women's Hospital/ Dana Farber Harvard Cancer Center/ Harvard Medical School in Boston, being involved in TMA experiments for the study of prostate cancer progression. This work was funded by the National Cancer Institute of the NIH (U.S.A.) Prostate Cancer S.P.O.R.E. program.

Chapter 2

2. State of the Art

To facilitate the *in situ* investigation of molecular markers in tumor samples, the tissue microarray (TMA) technique was recently developed [12]. TMA technology allows researchers to assemble hundreds of tumor samples into a single paraffin block. Sampling has been a major concern regarding TMA technology. How well do tissue samples, often with a diameter of 0.6mm, represent the entire tumor? Many of the initial TMA studies addressed this issue, by estimating the number of core replicates necessary and sufficient to achieve same results that would be obtained with conventional glass slides [24]. Another approach is to estimate the number of required cores to have evidence of well known association of a protein with some histological or clinical variables [25][26]. Some authors [27] also report that the sampling issue can be ignored and one core for each patient is sufficient for almost all TMA based study purposes. These different approaches suggest that there is no unique answer to the question of sampling. Perhaps the most critical deciding factor depends on tissue homogeneity and biomarker characteristic.
There have been also some attempts to formally compare the consistency of TMA analysis from one center to another. However, only a few of these have been published to date [28].

In the last few years an increasing number of studies has been published based on TMA technique, mostly aimed to investigate single protein expressions [29] to assess diagnostic or prognostic marker capability. TMA have recently been used to analyze panels of proteins trying to characterize specific disease molecular profiles [9].

Even more recently, TMA based investigations are included in cross platform approaches [30][31].

The focus of this chapter is not to review all of the biologic studies associated with TMA, but rather to focus on those research studies aimed at managing TMA data with respect to data collection, data organization, data analysis and data sharing.

In the setting of high throughput experiments where considerable amounts of heterogeneous data are present and many parameters are involved, the lack of appropriate tools to handle experiments and data represents an important problem for the future of the TMA field with respect to data interpretations. Moreover, integrated and comprehensive solutions would work for emerging studies, which could share previously studied cohorts or could include previous detected information and knowledge, taking advantage of well organized databases.
Another important feature of TMA based studies is the growing need to share data and information among different institutions (e.g., multi-center studies, clinical trials, etc.). In such a setting, the implementation of standard data exchange protocols becomes critical as up to now there have not been standard approaches to collecting data at different institutions and sometimes even within the same institution and as centralized solutions are not feasible. As a result of several TMA workshops in the area of TMA bioinformatics, some authors have reported [23] on a community-based, open source tool for sharing tissue microarray data; the exchange specification approach is based on a well-formed XML document. Others [32] proposed a specification of semantic metadata schematics for TMA in a peer-to-peer infrastructure design.

As for data handling, the use of a large spreadsheet has been the standard solution. This approach is useful for experiments on a one-time basis, but becomes very cumbersome when analyzing multiple markers on a given specimen or when having multiple observers render diagnoses and scores on a given specimen [33].

A grid sheet for scoring and analysis and a spreadsheet containing the list, type and position of each tissue core are, for example, provided by the Tissue Array Research Program. It is a collaborative effort between The National Cancer Institute and The National Human Genome Research Institute (http://ccr.nci.nih.gov/tech_initiatives/tarp/), which primary objective is to develop and disseminate Multi-Tumor Tissue Microarray slides and the related technology to the cancer research investigators.

Few papers have been published so far on TMA data organization and management. Liu *et al*. [34] presented a system for high-throughput analysis and storage of TMA immuno-staining data, using a combination of commercially available software and novel software. Similarly, Shaknovich et al. [35] proposed a way to manipulate TMA data and images, using commercially available software. Manley *et al*. [36] highlighted the peculiarities related to high density information glass slides, providing a source of inconsistency between tissue core sections and donor blocks identifiers. They proposed the used of a relational database for the better organization of TMA data. A validation study of the prototype system was then published [28].
Recently a web-based prototype for imaging, analyzing and archiving TMAs was proposed [37], mostly focusing on automatic evaluation of biomarkers. It also faces the problem of automatically extract single tissue locations to allow unsupervised registration of arrays, proposing an approach which works well with rigid rotations of the core section array on the glass slide.
Academic groups are working on the development of systems integrating commercially available software for the acquisition of digital images and the automatic evaluation of markers with custom solutions of data organization and management.
For instance, the TMA Profiler system developed in Rubin laboratory (http://rubinlab.tch.harvard.edu/htma/profiler/index.jsp) integrates the Chromavision system outputs (Chromavision Medical Systems, Inc., San Juan Capistrano, CA. Automated Cellular Im-

aging System (ACIS II)) in a web based platform to handle TMA experiment data (see additional data in [92]).
Similarly the Johns Hopkins TMA Laboratory, (http://tmalab.jhmi.edu/) employs an open source software (for academic use) to manage a TMA Database, TMAJ [38]. It allows the storage of a wide variety of information related to TMA samples, including patient clinical data, specimens, donor blocks, core, and recipient block information. A dynamic database structure allows users to add custom fields for different organ systems. The client application facilitates automated and manual entry of data related to patients, specimens, tissue blocks, and tissue subblocks (individual pathological diagnoses). The system allows users to design their own. Digital images generated by the Bacus Labs Inc. Slide Scanner (BacusTM laboratories, Lombard, IL, http://www.bacuslabs.com/) are imported into the database and available for on line visualization and evaluation.
The AQUA system [57] uses a custom imaging microscope system for scanning TMA slides stained with fluorescent markers. It uses an object recognition approach to exactly identify the spatial coordinate of each spot, but lacks in ordering and assigning them to proper patient and/or clinical information based on construction information.
Some other systems aimed to automatically acquire and evaluate TMA samples are available, such as the TMALab ™ (Aperio technology**,** http://www.aperio.com) or the Pathfinder™ Morphoscan™ (http://www.imstar.fr/).
These solutions are appropriate to handle very good quality TMA slides, usually by superimposing a grid on the panoramic overview of the slide, but they require considerable manual intervention if the samples are distorted (for instance, asking the user to fit the edge of each grid cell appropriately in case of misalignments).

With respect to data processing, usually conventional approaches are taken in handling TMA data. The most common approaches are statistical and machine learning ones. Some TMA studies have attempted to address a few issues related to data analysis

(see chapter 5) such as dealing with dichotomization protein expression levels and pooling data from replicate tumor samples [39][40].
Usually replicate tumor samples are included in TMA dataset in order to account for tissue heterogeneity. The pooling of protein level measurements detected on replicate samples is commonly straightforward adopted, therefore neglecting part of the information. An approach to model data uncertainty is proposed by Bhattacharyya1 et al. [41]; they include the data uncertainty in classification and relevant feature identification algorithms based on robust sparse hyperplanes, by associating each data point with an ellipsoid parameterized by a center and covariance matrix. A different approach in handling the data uncertainty may regard the use of multilevel models (or hierarchical models) [42][43], suitable to analyze information available at different levels of observation units (as for example, in meta-analysis of separate randomized trials).

Chapter 3

3. Molecular profiling at protein expression

3.1. Scenario

Gene expression is the process by which genomic information is converted into the structures and functions of a cell, by producing proteins from DNA coding sequences (see Figure 2). Genomic expression is more and more under investigation, trying to depict molecular profiles characterizing specific disease states. New technologies, microarrays, allow tracking the translation of genome sequences into gene functions (functional genomics), leading to the identification of highly informative genes and pathways with potential impact on understanding disease development and progression [44], concurring to improve diagnostic and treatment modalities and to detect targets for new therapies [45]. Despite different kinds of investigation, all microarrays generate large datasets by simultaneous detections under the same experimental conditions. In particular DNA microarrays (DMAs) analyze mRNA expression, and Tissue microarrays (TMAs) evaluate DNA, RNA or protein targets through *in situ* investigations (analyses performed on tissues).

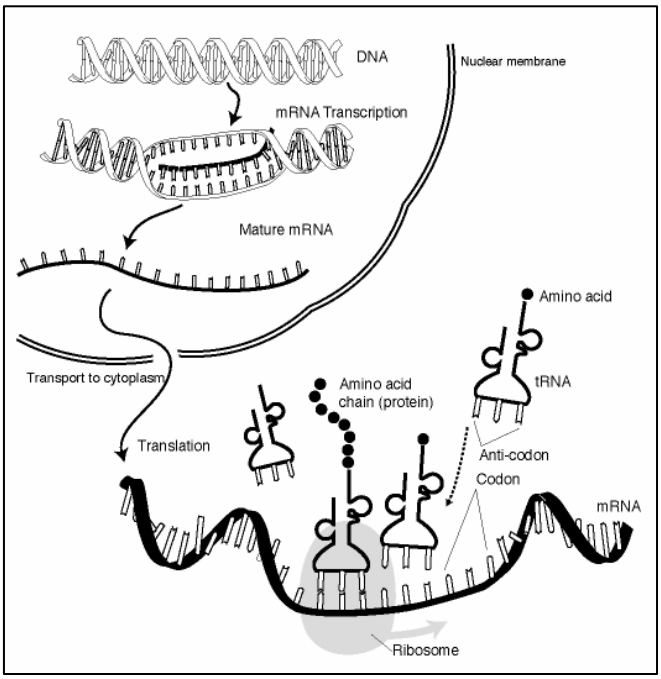


Figure 2 - Gene expression (Picture taken from National Human Genome Research Institute (2001)). [Gene expression is the process by which a gene's information is converted into the structures and functions of a cell. Gene expression is a multi-step process that begins with transcription and translation and is followed by folding, post-translational modification and targeting. The amount of protein that a cell expresses depends on the tissue, the developmental stage of the organism and the metabolic or physiologic state of the cell.]

Microarray technology is extremely powerful because of the high-throughput nature of the approach, giving great advantages with regards to experimental execution time and experimental homogeneity, both being particular relevant for comparative studies.

In molecular profiling studies of human cancer, integration of DMA and TMA experiments provides a powerful approach;

genes selected by DMA studies can then be simultaneously investigated by pathologists on large set of well characterized solid tumors by immuno-histochemical or nucleic acid hybridization techniques.

3.2. Data management problem

How can we make sense of all the data produced by these high throughput technologies? How can we be sure that our data are of good quality and therefore that data analysis results are reliable?

Common critical points of these new experimental techniques are data control and data quality, which should be account from the assessment of experimental data production and collection to data interpretation.

In TMA studies, data collection and organization are extremely relevant: heterogeneous data types are included in studies (gene expression levels, histopathological data, clinical data), several data sources are involved and the evaluation procedure of gene expression levels on TMA samples is definitely prone to error [45], as for instance association errors in designing the block array or evaluating a slide may easily occur.

Efficient retrieval of data is also desirable, both as TMA studies are usually based on several single experiments on the same cohort of patients and for new experiment design. Data and information sharing is also mandatory as often multi-center studies are performed.

Components to be considered in this context are thus i. critical assessment of experimental design and organization, ii. reliability assessment of experimental data and iii. technological support to organize and collect data, also implementing automation.

As far as data organization is assessed the problem shifts to data analysis and interpretation. Even if a lot of work has been done on gene expression data analysis and traditional approaches might be successfully applied very often, the TMA data analysis phase is not trivial. Great attention in preprocessing phase should be

adopted; appropriate analysis approach must be carried out depending on specific biological or clinical question. (i.e. hypothesis driven studies).

Even if technical support is provided and automation is employed to speed up the experimental process, a very important step is data preprocessing. Contrarily to what happens on cDNA microarrays where single elements are processed, tissue samples are part of solid tissue, thus being spatially heterogeneous. Every TMA glass sample, even if consecutively obtained, contains slightly different tissues (see section 5.3.1).
Another kind of heterogeneity is addressed by intra-tumor heterogeneity: protein expression may vary across the tumor. Multiple measures of protein expression on the same tumor must be handled during data analysis. The latter is a problem that is usually skipped in TMA data analysis, as conventional statistical and machine learning approaches do not account for data uncertainty.

To efficiently and practically face the above mentioned problems a strong interdisciplinary effort must be done. Accurate attention must be given to all the aspects concerning this experimental technique and its possible integration with experimental data obtained with other platforms.
Knowledge Discovery and Data Mining (KDDM) [46][47][48] processes provide an ideal framework for bioinformaticians to fit: from understanding the domain, forming a consolidated dataset from data sources, cleaning the data by selection and preprocessing, extracting regularities in the data and formulating knowledge in the form of patterns and rules.

Lack of integration and uncontrolled data collection activities not only affect study results, but also damage future activities sharing same patient or same data or protocols, which could have taken advantages of all previous work.

The next session describes TMA techniques with particular emphasis on common problems that can affect experimental conditions.

3.3. Tissue Microarray Technique

The first attempt to increase the throughput in terms of number of samples for simultaneous *in situ* analysis was proposed by Battifora [49] in 1986 with the 'tissue sausage' technique. The intent of this "tissue sausage" was to increase throughput in the evaluation of antibody staining properties for a variety of tissue types. The Tissue Microarray [12] technique improved on the haphazard arrangement of the tissue samples and by placing the samples in an ordered array allows for high-throughput *in situ* experiments using immunohistochemistry, FISH, or *in situ* hybridization. The TMA approach has a dramatic advantage over the conventional approach in performing *in situ* experiments on standard glass slides. The standard approach requires performing many experiments on separate slides with the major drawback of inter-experimental differences due to variability of staining from one slide to the next and an inefficient use of tissue samples. The TMA approach allows for the simultaneous staining of hundreds of tissue samples from as many patients. Thus this TMA approach is a high throughput process, which ensures experimental standardization across all patient samples (i.e., all tissue samples are pre-treated and treated under identical conditions). Perhaps one of the most critical advantages is that TMA technologies conserve the limited tissue resource, which is vital given the increasing number of candidate genes that need to be explored.
The TMA technique is demonstrated in schematic form in Figure 3. Tissues are arranged in a pre-ordered matrix. Cylindrical tissue biopsies are transferred with a biopsy needle from carefully selected morphologically representative areas of original paraffin blocks (donor blocks). A hematoxylin and eosin (HE) stained section of the donor block is usually aligned to better locate the representative areas. Core tissue biopsies are then arrayed into a new

“recipient” paraffin block by using a manual or an automated tissue arrayer (e.g. Manual Arrayer, Manual Arrayer, Beecher Instruments Inc, Sun Prairie, WI, USA), using a precise spacing pattern along x and y axis, which generates a regular matrix of cores. TMA topology design is not unique and may vary depending on the study design, number of tissue cores, and the types of potential experiments.

A TMA block (block array) can theoretically contain over 1000 cores but for practical reasons, most experts in the field do not place more than 600 hundred tissue biopsies on a single block. The number of samples may also vary depending on the needle diameter used to transfer the samples (from 0.6 to 2 mm in diameter) [12][45]. TMA sections are serially obtained at 4-5 micrometers thickness [50] with a microtome, a precision instrument designed to cut uniformly thin sections of materials for microscopic examination purposes. About 150 sections can typically be obtained from one block array. However, the number of sections will ultimately depend on the depth of the tissues originally placed into the recipient block. [51]. Construction of identical TMA blocks (settled up with the same donor materials and in the same order) is also common, to further scale up the method.

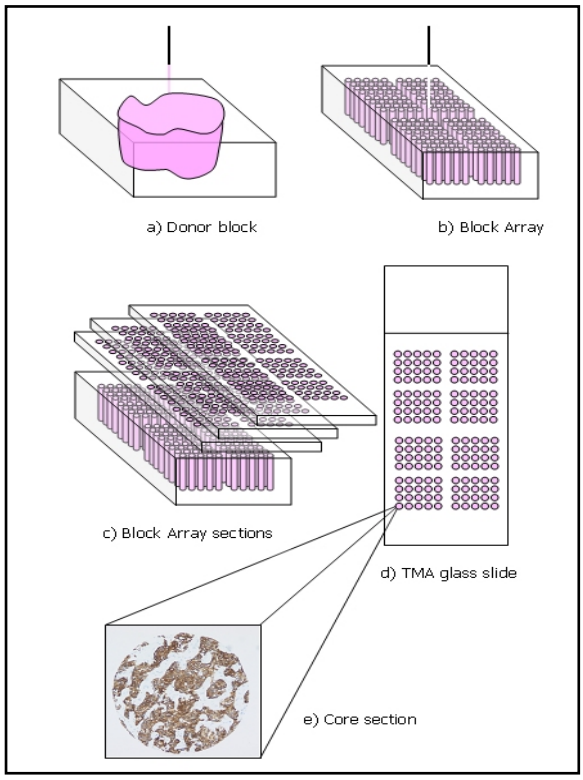


Figure 3 - Tissue Microarray Technique scheme.

Sampling is always a concern in the field of pathology. Rarely do pathologists review an entire organ but instead take representative samples. TMA technology is no exception to this common practice. How well these tissue samples, often as small as 0.6 mm in diameter, represent of entire tumors has been the focus of several recent studies [52][53][54]. Some authors [55] also report that because TMAs are mostly intended for population level screening tool than for diagnoses on individual cases, this aspect is not critical. The results of those studies are dependant on tumor types and study purposes. A biomarker with homogenous expression

throughout the entire tumor will not require as many samples as a biomarker that is only focally expressed by the target tissue. In order to capture some of the intra-tumor heterogeneity, multiple cores are often taken from each donor block and arranged in the same block array.

3.3.1 Common problems with TMA section preparation

During the preparation of TMA slides, due to technical problems the content and/or the perfect alignment of the array may be altered. For example, as the donor tissues may not all have the same depth, as one cuts deeper into the TMA, some individual TMA cores will disappear due to the absence of that particular tissue core (i.e., exhausted tissue). Core sections may be fragmented during the cutting process with even the most sensitive microtome. The transfer of the TMA section from the microtome onto the glass slide may also lead to tissue loss. Through imprecise processing, the array may be distorted and one or more rows and/or columns may be bent, folded or moved from their original position (see Figure 14).

To reduce the array distortion and bending problems, the TMA sections may be placed on the slides using a tape transferring system (Instrumedics, Hackensack, NJ). With this system the sections are captured flat and uncompressed on a tape-window as they are cut and then transferred to an adhesive coated slide. This procedure, adopted by many laboratories, requires altering slightly the staining protocol. The tape system reduces distortions and bending problems, but, with varying degrees of frequency, misalignments are almost always present even on the highest quality TMA samples.

Due to tissue heterogeneity, the original targeted tissue placed into the TMA may appear different as one goes deeper into the TMA block. This is expected given the 3 dimensional nature of the samples being used. This limitation requires that the samples be evaluated repeatedly at all layers.

3.3.2 *In situ* investigation

All *in situ* experiments that can be performed on standard tissue sections can also be performed using TMAs. The TMA approach has the added advantage, however, of being able to increase throughput and limit experimental variability often encountered with standard slides. TMAs can be used to detect DNA, RNA or protein expression, by using *in situ* hybridization or immunohistochemistry[56].These methods do not alter the tissue morphology and therefore information on distribution of the expression within the tissue can be obtained at the level of individual cells, even in cases where the expression occurs only in a few cells. One major advantage of examining expression *in situ* is the ability to determine the exact location of expression down to subcellular level. For example, expression of a biomarker such as beta-catinin in colorectal cancer has a dramatically different meaning whether it is expressed in the cell membrane (inactive form) or in the nucleus (active form) [57]. Expression detection and localization concur to potentially interesting information.

The expression evaluations of these *in situ* tests have been traditionally performed by pathologists using microscopy (either bright or dark field). For example, immunohistochemical experiments are often evaluated by pathologist for staining intensity using a four nominal point scale (negative, weak positive, moderate positive, strong positive), and the percentage of the stained area. However, the human eye, even if trained, cannot detect subtle differences in staining intensity on a continuous scale, in particular at very low and very high levels of the scale. Moreover it is well known that nominal categories are subjective and inter- intra-observer agreement is usually poor. In contrast, automatic detection of immunostaining provides continuous quantitative measures, also ensuring reproducibility due to objectiveness of the measurements (see chapter 5.2). More recent work in the field of TMA biomarker development has focused on the use of semi- or fully-automated image processing. The continuous quantitative measurements obtained by automatic evaluation can better detect

the information of the investigated target, allowing the detection of subsets of tumors not seen using human/pathologist based assessments as demonstrated in [57], where fluorescent probes were used.

3.3.2.1. Immunohistochemistry

Immunohistochemistry is the process of detecting protein expression *in situ* using antibodies targeted at the protein of interest. The antibodies used can be polyclonal or monoclonal in origin, the monoclonal ones being more specific in nature. Immunohistochemistry is widely used in the evaluation and diagnosis of cancers and other disease states. Therefore, most practicing pathologists are familiar with this technique. Regardless their source (monoclonal, polyclonal or recombinant) not all antibodies work equally well. Specificity, sensitivity and antigen preservation must be considered [56]. False positives can be due to cross-reactivity with related or unrelated epitopes. False negatives can arise when the antibody fails to recognize the epitope within the tissue, the epitope was altered by fixation, the epitope is present at too low level or it is inaccessible for the antibody due to other reasons (protein-protein interactions, cross-linking or modification).

Pre-processing technique may also play an important role in the performance of these *in situ* techniques. Tissues are routinely placed into a fixative such as a 10% formalin solution. After a variable time in this fixative, the samples are embedded in paraffin blocks. This procedure is used to preserve tissue morphology and stabilizing the tissue from degradation, but a drawback is that it can modify target antigens. For routine purposes, probably 99% of all processed tissues reside in this state. An alternative, most often associated with research protocols or molecular analysis, is the procedure where samples are snap frozen requiring no fixative such as formalin. Research investigators are developing arrayers

capable of handling frozen tissue samples instead of paraffin-embedded ones [39][58].

Chapter 3

Chapter 4

4. The TMA data management system

Before construction of our TMA management system, we took into consideration some of the critical issues related to the standard needs and requirements of TMA experiments:

- TMA studies may often involve considerable numbers of patients, heterogeneous data are involved, and different kind of users interacts at different times. Users are often from different institutions;
- TMA technology is prone to several common types of errors: association errors in designing the block array or evaluating a slide may easily occur;
- Patients included in a TMA study (included in one block array) might be later on included in some other TMA studies or tissue based studies. (One to many type of relationships need to be considered).

These observations led us to develop a system with the following features:

i. patient centered and not experiment centered,
ii. work flow oriented accounting for all phases of TMA experiments (patient data collection, design experiment, data production),
iii. optimal level of automation,
iv. web based to allow for easy access and strong inter-institutional capabilities.

Based on these guidelines we constructed a system, where medium-long term advantages are data sharing, data quality enhancement and data reusability. We paid attention to balance flexibility requirements and constraints in user interface implementations, in accounting for the storage of all the details that

make an experiment reproducible (for ex. staining preparation protocols, digital image acquisition setting as lamp intensity, filters) and speeding up data acquisition and evaluations implementing automation.

4.1. User Requirements, Need Assessment and System Development

Need assessment and user requirements have been defined, collected and analyzed from the very beginning of the work/project. We conducted face-to-face interviews with the different potential system users, focusing on workflow definitions and working constraints.
We took advantage of an already available common working language between the potential system users and our group, due to previous collaborations. The main potential users of the system are pathologists, biologists, laboratory technicians and clinicians.
Modifications of system features were adopted not only during the course of the system design phase but also during the implementation phase and test phase, consistently with generic life cycle models [59].
We implemented a first prototype running in local environment. During a four month period of usage, we collected comments and observations on data flow and interface acceptance. Modifications were then implemented in the web version of the system.

4.2. TMABoost System

The TMABoost system [60] consists of i. a web based application to collect all data involved in TMA studies: patient and tumor data, donor block information, block array information, array slide staining, core section evaluations, etc., ii. a relational database to store the data and iii. a digital TMA environment based on the usage of a robotic microscope and image processing algorithms to exploit digital pathology [61][62]. This environment automatically acquires glass slide overview images and single

core section images. Images are stored in the database and can be visualized by pathologists through the web interface. Moreover, automatic quantification of biomarker expression can be computed on these digital images, avoiding subjectivity in human evaluation and allowing novel associations thanks to the continuous nature of produced data.
The choice of developing a web-based system was driven by different motivations. First, a web-based system facilitates the collection of data from several institutions, providing access from virtually everywhere with an Internet connection. As studies between multi-centers is becoming more and more common, this system needed to be capable of gathering material, information and clinical data for specific research purposes from multiple sites. This web-based system also allows pathologists from these different institutions to evaluate the same TMA images.

In the following sections the system will be presented in terms of system architecture and system components separately. This allows describing more intuitively system functionalities, data input, data retrieval and automated data collection. In particular we describe i. the system core, composed by a relational database, a middle/processing layer, and a web browser interface, and ii. the Digital TMA Environment.

4.2.1 System Architecture

Due to the functional requirements the general architecture of the system has been designed following standard web-based 3-tiers architecture [59].
System modules are logically grouped into three layers (see Figure 4) for which a brief description is here reported.
The presentation layer is the representation of the system state to the user. It is made of two modules: the web browser and the Digital TMA Environment. Web pages are mostly Dynamic Hyper Text Markup Language (HTML) pages. We used Javascript language to provide several controls on user data entry. For pages

that require a lot of data entry, we used eXtended Markup Language (XML), together with Microsoft XML DOM, both to organize and represent the information. This solution, coupled with Javascript code, allows performing some computation on the client side, reducing server workload and client/server communication. We used Microsoft Internet Explorer 6.0 as web browser.
The Digital TMA Environment interacts with the database through the Web System Communication Module (see Figure 11). This module establishes a connection with the web server to retrieve relevant information from the database: the block array map and the block array construction parameters, information required for the object recognition procedure. Moreover, the Web System Communication Module returns to the server both the acquired digital images and the results of the automatic biomarker evaluation. Digital images are sent as an independent background process. This allows the user to perform other tasks, while transferring the images. The overall activities of this module are performed seamlessly to the user.

The middle layer is made up by the web server; we used Microsoft Internet Information Server 5.0. We employed Active Server Page (ASP) and Vbscript/Javascript on the server side to allow and control the communication between the user and the database. In addition the source code of the middle layer is written on a central server and thus system upgrades and improvements are easily performed and immediately available to all users.
A Secure Socket Layer (SSL) has been established to ensure data encryption between the client (user workstation) and the server, preserving security and confidentiality of communications. Moreover, authentication of the user is based on login and password that, thanks to SSL, are properly transferred to the central web server in a secure way.

The Digital TMA Environment can send images to the middle layer through standard HTTPS protocol. The middle layer then

stores them into the proper table of the database as Binary Large Objects (BLOB).
In the data layer we employed a relational database (Microsoft SQL Server) to store the data. SQL server is a XML enabled Database management System (DBMS) [63]. The database is implemented behind a firewall.

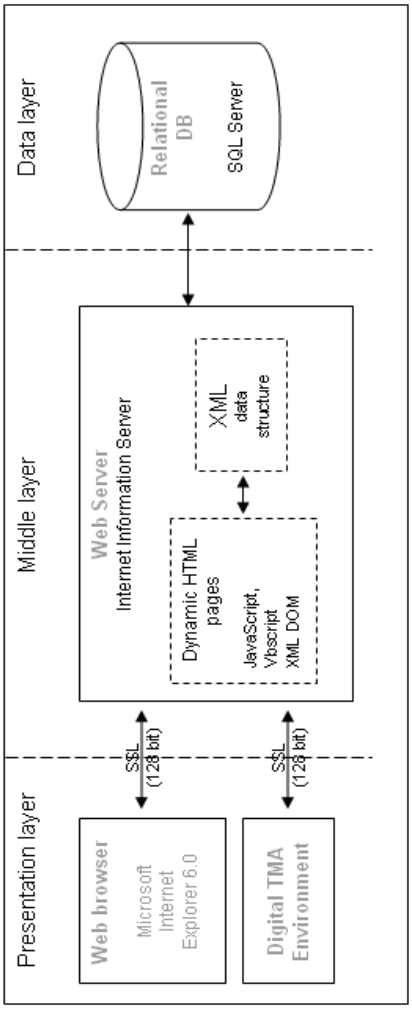


Figure 4 - Web-based 3-tiers architecture of TMABoost system.

4.2.2 System Components

4.2.2.1. System Core: the database

A relational database [64][65] addresses data organization requirements in an efficient manner, enabling data consistency, correctness, and completeness. Unique identifiers - automatically computed by the database engine - are used to link records between different tables, giving considerable advantages in terms of performance. This solution allows avoiding inconsistencies in the data and, together with the middle layer, reduces possible errors in linking biological/genetic, pathological, and clinical information. The database management system we employed is Microscoft SQL Server 2000.
The relation diagram of the main tables is shown in Figure 5.
Tables can be grouped as i. patient and tumor content tables (tb_paziente, tb_tumore_paz, tb_tu_metastasi, tb_tu_recidiva, tb_donor_block), ii. TMA experiment preparation tables (tb_block_array, tb_block_array_map, tb_array_slide, tb_core_section, tb_acquisizione_slide, tb_images) and iii. TMA experiment data tables (tb_valut_man_core_section, tb_valut_auto_core_section, tb_images_analysis).
In addition, there are nine auxiliary tables, mostly of them independent from TMA technique, containing system user information, tissue provenience information, histology classification, biomarker descriptions, laboratory protocol codes for block array and slides preparation, etc..

A one-to-many relation between the tumor table and the donor block table does not allow inserting donor block information, usable for a TMA block design, without previously inserted tumour data.
Similarly, it is worth noting the relation between the donor block table and the block array map table: it ensures consistency between tumour/patient data and the cores of the block array. This relation is crucial, as an error in the block array map propagates

along all experiments based on that block. The block array map is unique for each block array and every successive data related to one TMA glass slide relies on the map.

The system allows each core section be evaluated multiple times, both by humans (tb_valut_man_core_section) and by automatic procedures (tb_valut_auto_core_section). The one-to-many relation between the array slide table and the acquisition slide table accounts for multiple acquisitions of the same TMA slide, for instance under different wavelength filter conditions, different lamp intensities, different field filters, etc..

Individual patient identifiers are not included in the database, which contains only de-identified data [48]. The system automatically assigns a unique code to each patient.

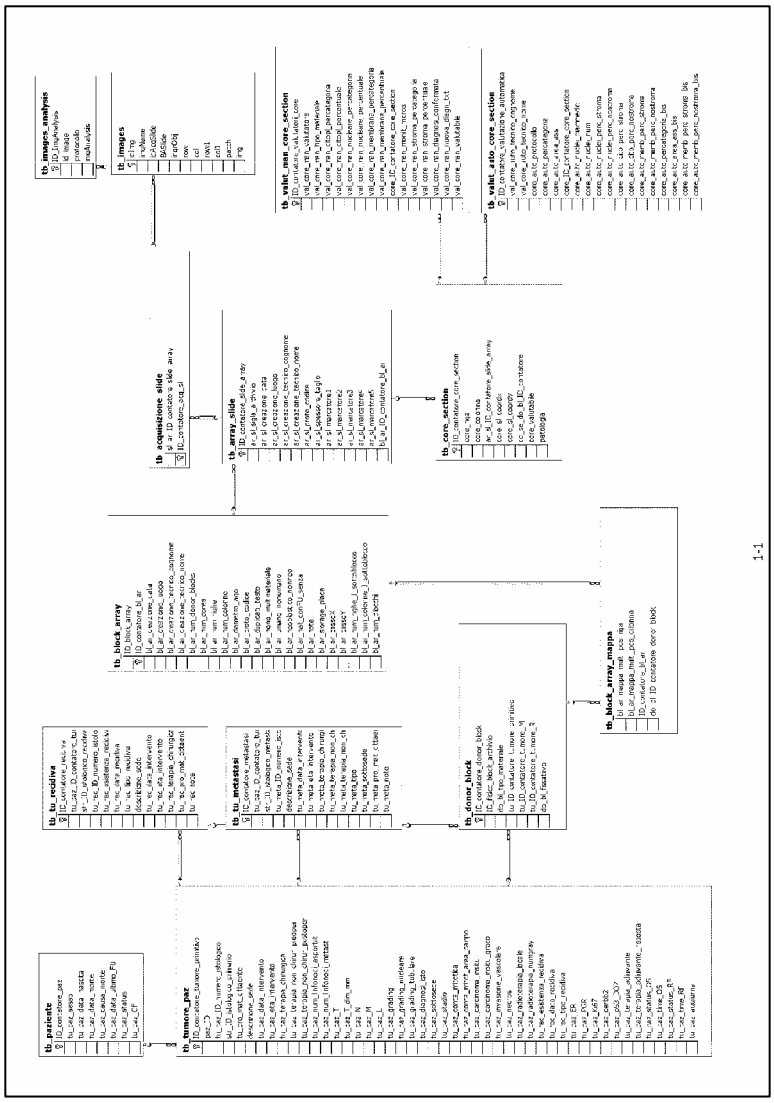


Figure 5 – Relational diagram of the TMABoost database.

Database Roles/User profiles

The DBMS manages user authentications by Login and Password. The DBMS also manages user roles. Seven profiles have been defined, granting role-based access to system data; each profile differently allows operations on table records as *select*, *insert*, *update*, and *delete*.

Each user is associated to one or more profile.

The defined profiles are the following: technician, clinician, evaluator, pathologist, automatic evaluator, administrator, demonstration user.

All the profiles have been granted read access to the database tables and, with exception of the administrator, none can delete records.

The technician role has write access on the tables concerning the donor block preparation, the preparation of the block arrays, the maps, and the staining of the glass slides. The clinician role has write access to the tables concerning the patient and the tumour. The evaluator role can insert data only in the core section evaluation table. The pathologist role has the broadest entry access to the database: it embraces all the permissions of previous users.

The automatic evaluator profile is the one used by the Digital TMA Environment to access the database: it can insert data in the core section table and in the automatic evaluation table.

The demo user profile has been added to allow potential new users to navigate into the system for a finite time period. The middle/processing layer grants their access to a test version of the database.

Constraints to handle data ownership are not yet implemented. They will turn out to be useful to allow individual research group to share or separate their data, based on study and sample permissions. These permissions can in turn be used to securely limit access to specific specimens, blocks, array-blocks, and sessions.

Database Triggers

SQL Server supports using triggers as a kind of stored procedure. Triggers are commonly executed when a specified data modification (for instance an attempt to delete a row) is attempted on the table on which the trigger is defined. We implemented some triggers to seamlessly perform control or updating operations on data. For instance, any change in patient's clinical data, such as status (alive with relapse, alive without relapse, dead for disease, dead for other causes), last follow up date, etc., automatically updates related information for survival analysis studies. Triggers can, on demand, be implemented on additional fields to address different endpoints. For instance, PSA failure in prostate cancer might have multiple definitions.
The usage of triggers is relevant, since clinical information can be repeatedly updated by different users.

4.2.2.2. System Core: the middle/processing layer

Besides allowing the communication between the user and the database, the middle/processing layer performs several processing steps to control data workflow. It provides the user with the proper interfaces for collecting and retrieving data, as better described in the following sections. Moreover, it controls user input performing a first filtering level on inserted data.
It also handles the communication with the database, by saving and requesting data. A fundamental aspect performed by the middle/processing layer is mapping the relational structure of the database into a hierarchical one by using eXtended Markup Language (XML). This solution allows exchanging data back and forth between the client and the server in a more efficient manner, reducing communication workload, and performing part of the computation on the client side. For instance, when the user is evaluating a slide, all core section information is downloaded once from the database as a XML structure. When the user selects a certain core section, only the corresponding data are shown

through JavaScript code executed at client side (see Figure 6). The user then inserts or modifies core section evaluations through the HTML interface. Seamlessly to the users, these changes update the XML structure on the client side, by JavaScript code. Only when data are completely inserted or upon specific user request the XML structure is sent back to the web server. The middle/processing layer then processes the XML structure and extracts the data in a suitable format to update the database.
Same approach can be implemented to extract data accordingly to every XML predefined structure, as the one proposed by [23] to exchange TMA data.
A key point of the middle/processing layer is the communication with the Digital TMA Environment (see Section 4.2.2.4). In particular, the exchanged information regards array slides and block array maps, automatic biomarker evaluation data, and digital images.

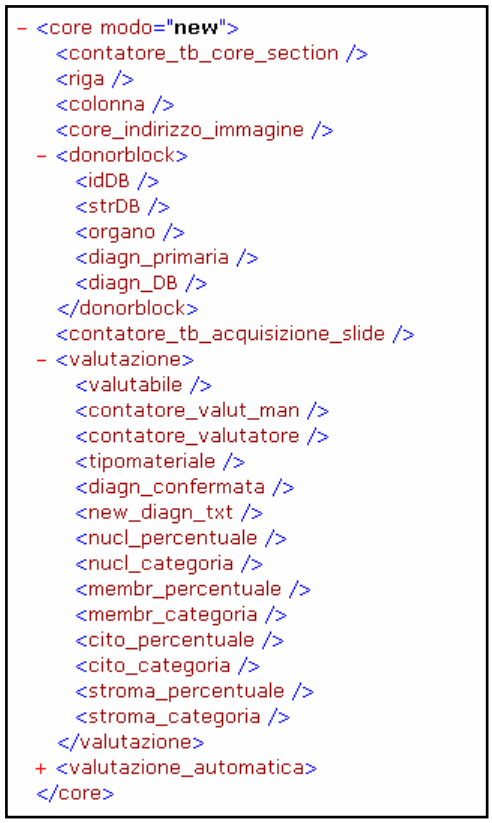


Figure 6 – XML extract: single spot description data and staining evaluation data.

4.2.2.3. System Core: the web browser interface

Particular attention has been given to the user interface design, addressing different needs of different users that operate in distinct work phases; workflow needs, differentiated for input and retrieval phases have been considered.
The web system interface is divided in to two frames: a navigation frame and a main frame. The navigation frame (see left frame

in Figure 7) allows the user to select different levels of information, as patient identification number, tumor and donor block data, block array or array slide data. For each group of information, inserting and retrieving pages are available in the main frame. Searching and retrieving data can be done through interfaces that present the user with predefined criteria for specific queries (dynamic queries). Results are then reported as lists of tuples, sorted as requested. Result lists can be used to further navigate in to the system.

Different types of constraints are made on web interface level to increase data reliability. Constraints on most of the data insertion fields are employed. Controls on data type entries (date, string, numbers) are present, control boxes for list choices are implemented as far as possible (pick up tables). In some cases the list items depend on other selection. For example, during tumor data insertion as the organ is selected, the appropriate lists for T, N and M (Tumor-Node-Metastasis classification) are loaded.
Accordingly to database constraints, some data fields are compulsory (controls made when the user tries to save, before sending the data to the database) and some others allow for data entry only accordingly to database contents. The latter case occurs for the block array map insertion: to insert data of each new punch, a list with all the available donor block codes is presented to the user.

One specific example of data insertion capability would be how the system helps guide the technician/user through the design phase of virtually constructing a TMA. This procedure is carefully defined i. to allow the user to select donor blocks to be associated to each cell from a list (no typing allowed), ii. to force the user to the pre-declared structure of the block array. To further help the technicians in the insertion phase of the block array map, some tools were implemented, such as an additional button to automatically insert data of multiple biopsies from the same donor block without reselecting it from the list and a graphical solu-

tion based on differentially colored map cells to visually identify such replicates (see Figure 7). This structure is highly preferred because block array map insertion is definitely prone to error. Eventually information about cuts/slices of each block array can be stored in the system at any time. At each new cut of a block array the system automatically assigns a label to it, composed by the code of the block array and an incremental number.

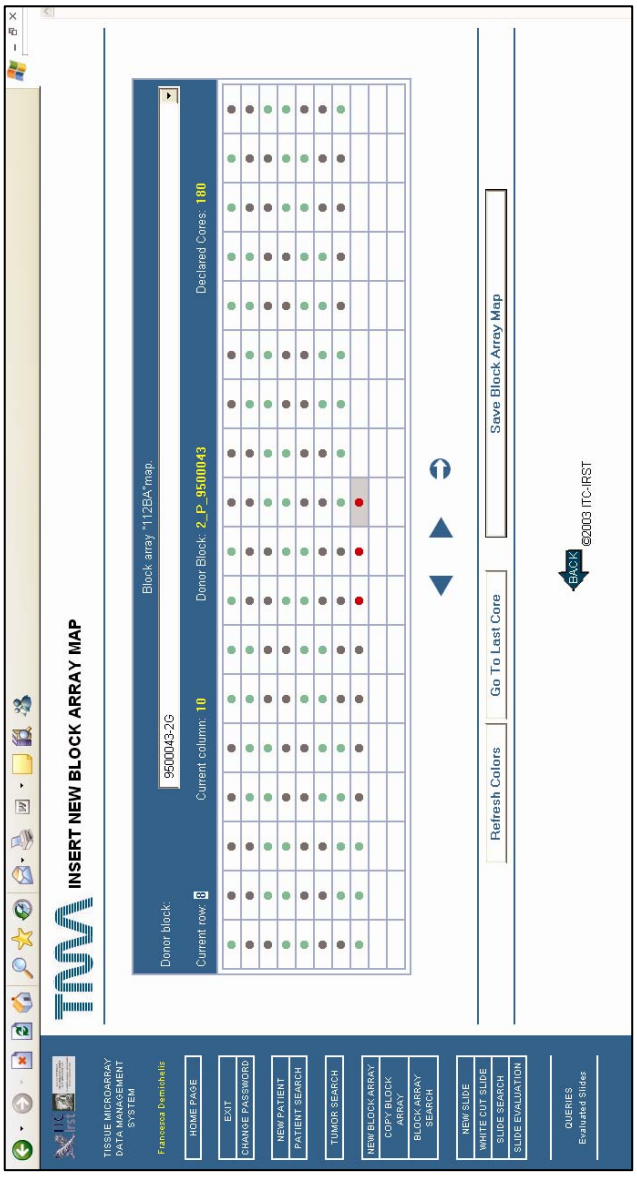


Figure 7 – Screenshot of TMABoost: block array map data entry.

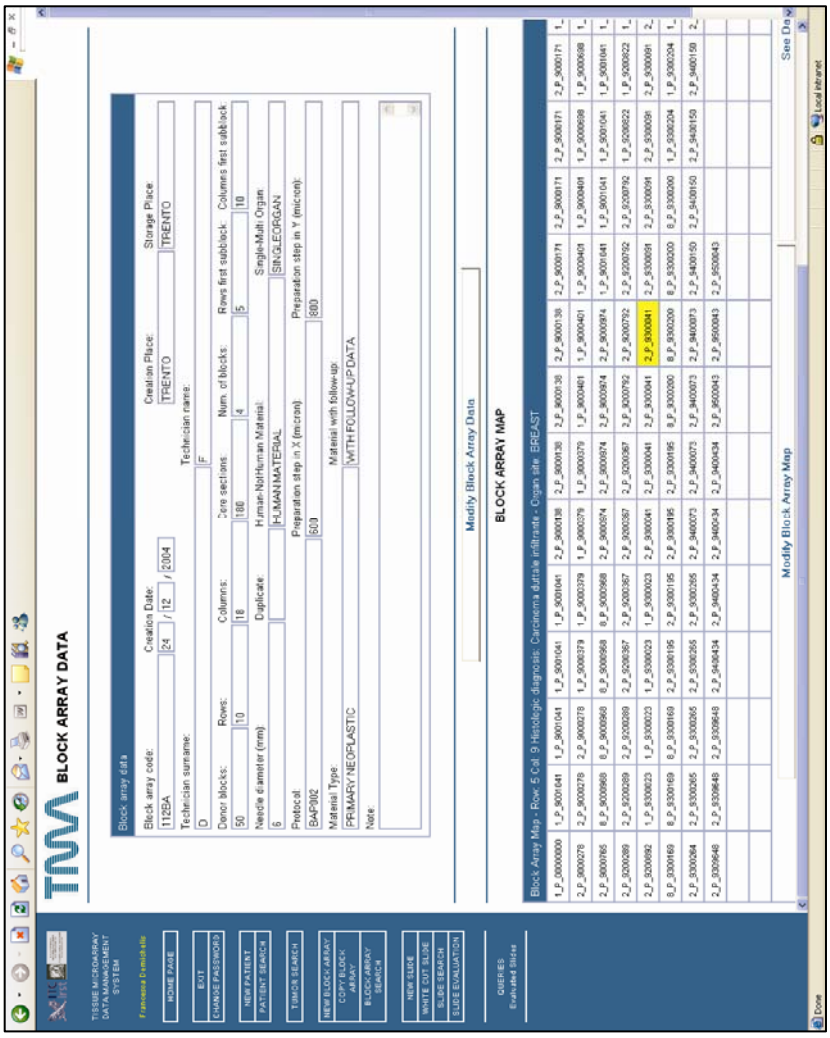
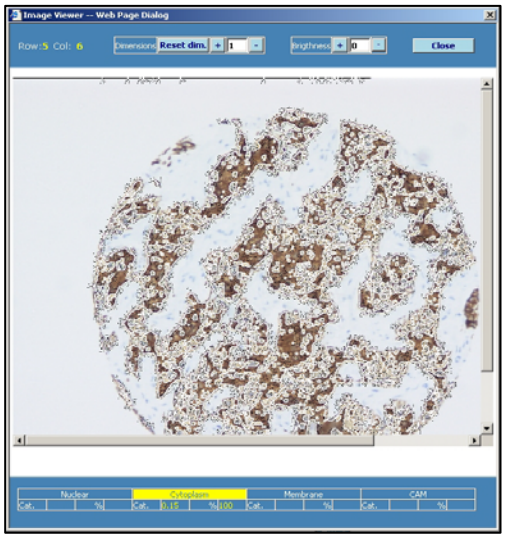


Figure 8 – Screenshot of TMABoost. Information of single block array made of 10 rows and 18 columns; preparation data and map of a block array is represented: as the mouse is moved over the map cells, data of the selected core appears on the top of the map.

Another example of the web interface is given in Figure 9. It shows the system interface for the input and the retrieval of biomarker evaluation of a single glass slide. Mouse clicks or shortcut keys can be used to move across the slide map, which is automatically prepared on the basis of corresponding block array stores data. In the top frame two types of information related to the selected core section are shown, i. data of the donor block where the core section comes from (tissue type, diagnosis), and ii. core section evaluation data (possible evaluation of the spot, diagnosis confirmation, biomarker evaluation data). The user may decide to visualize single core section images during slide evaluation or visualization, moving from one core section to another. Shortcut keys have been implemented to ensure fast data insertion.
Multiple evaluations of the same TMA slide are feasible and each is designated as a separate session. Each user (associated to the evaluator or pathologist profile) can edit only its own evaluations.



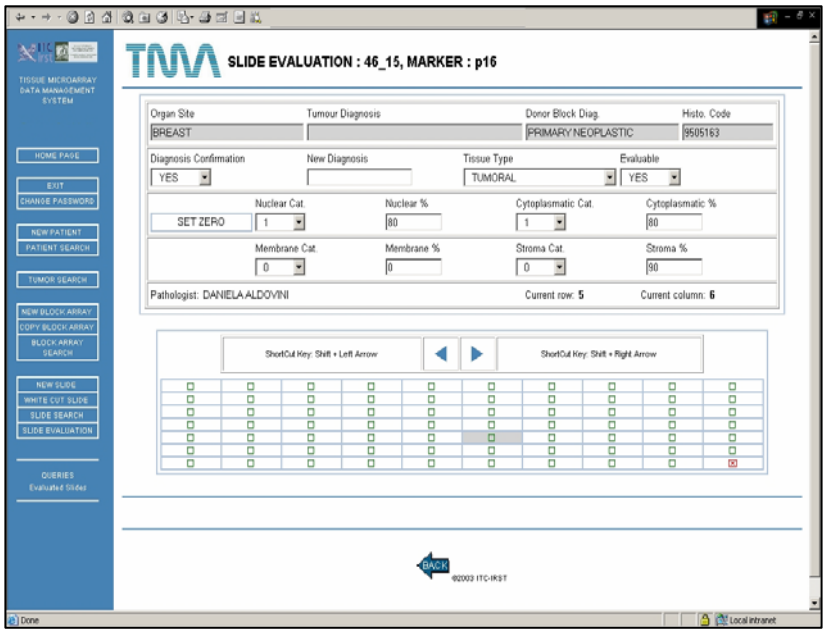


Figure 9 - Biomarker evaluation: TMABoost interface; the upper window shows the image of the selected core section.

4.2.2.4. Digital TMA Environment/ TMA Acquisition Environment

To ensure comprehensive information storage and large scale analysis, digital images of glass slides and of single core section must be acquired and stored.

The storage of digital images of each core section has two main advantages:

- it makes tissues available to pathologists involved in specific studies through the web, enabling on screen core section evaluation or review;

- it allows the automatic evaluation of gene and protein expression through image processing routines, avoiding subjectivity (see Figure 12).

Provided digital images are almost mandatory at least to offer complete information storing of TMA experiments, an efficient solution must be found to speed up this phase.
In addition, future changes in editorial policy may require submitting all images associated with TMA studies at time of review and publication, as is common practice with expression array studies.

A key point of the present system is given by the integration of an automated digital TMA acquisition system, 'connected' to the TMA database through the web server.

The TMA acquisition system employs a robotic microscope, which is a microscope drivable by software in each movement: stage movement on x and y axes, magnification change, z-axis movement, auto focus, etc. Robotic microscopes are largely used for telepathology applications, for the creation of digital slides (entire histological or cytological specimens) for several purposes [62] and recently for the acquisition of TMA slides.
An interface for a robotic microscope was already designed by our group to handle so called digital cases, sets of digital images representing whole glass slides.
Heavy modifications on the previous work were done to face TMA sample peculiarities. See Figure 10.
Comparing to conventional digital slide, the pedant acquisition of entire slide section is useless, as information (tissue) is not continuously spread on the glass, rather spottily localized. At the same time, checking on the fly if an image contains or does not contain tissue is not an efficient solution; in fact, depending on camera target and on magnification it might happen that, pedantically acquiring the slide, a portion of tissue is always present. Therefore a different approach should be applied.
Another aspect to be considered is the need to correctly associate each core section and, therefore the corresponding digital image,

to the proper donor tumor [37] that is a crucial issue in managing TMA slides.
As described in section 3.3.1, misaligned tissue samples due to processing errors or tissue loss occur with varying degrees of frequency but are almost always present even on the highest quality TMAs. Therefore the straightforward solution of applying a blind grid based acquisition is often not suitable and a non trivial solution for the automated core section assignment procedure is desired.
To address these aims, we implemented an image processing routine and an object recognition algorithm, working on an overview image of the entire slide. Each image can be automatically identified and acquired and, thanks to the information stored with the block array map in the database, associated to the proper tumor and the proper donor block.
We tested the accuracy of the routines on a set of TMA panoramic images. The object recognition algorithm we implemented revealed 96.8% of accuracy (5688 core sections out of 5878) (see 0).
This solution to automatically perform grid location assignment is particularly relevant and crucial; it allows high-throughput screening of TMAs speeding up the process and enhancing data quality related on the exact assignment of each tissue to original tumor.
The Digital TMA Environment component also integrates image analysis routines to automatically evaluate protein expression on TMA tissue samples.

The architecture of the Digital TMA Environment is reported in Figure 11. This custom made environment has been developed in Visual C++ in a modular way, to ensure re-usability of software components. Modularity allows adding, upgrading, and substituting components; for example a new camera could be integrated in the acquisition system, an image analysis routine could be upgraded or a new one could be added. In particular, the DM LA Controller module can be substituted to interface different robotic

microscopy equipments, letting the system being hardware independent. The two systems in use in Trento (Bioinformatics Group, ITC-irst and Santa Chiara Hospital, Pathology Department) employ Leica DM LA Microscopes (Leica Microsystem, Wetzlar, Germany) coupled with 3-CCD Sony DXC 390P video cameras (Sony Inc, Tokyo, Japan). Matrox Meteor II MC acquisition boards (Matrox Electronic Systems Ltd, Dorval, Canada) digitize analog video signal. Software applications are executed on Personal Computer Pentium 4 processors - 1.6 GHz - with Microsoft Windows 2000 operating system. Leica DM SDK 4.1.4 and Leica Video Autofocus libraries were used.

The processing routine (segmentation plus object recognition procedure) was developed by applying image processing and analysis operators contained in Aphelion 3.1e I mage Processing Library (ADCIS S.A., Hérouville Saint-Clair, France), and integrated in the Digital TMA Environment as ActiveX controls.

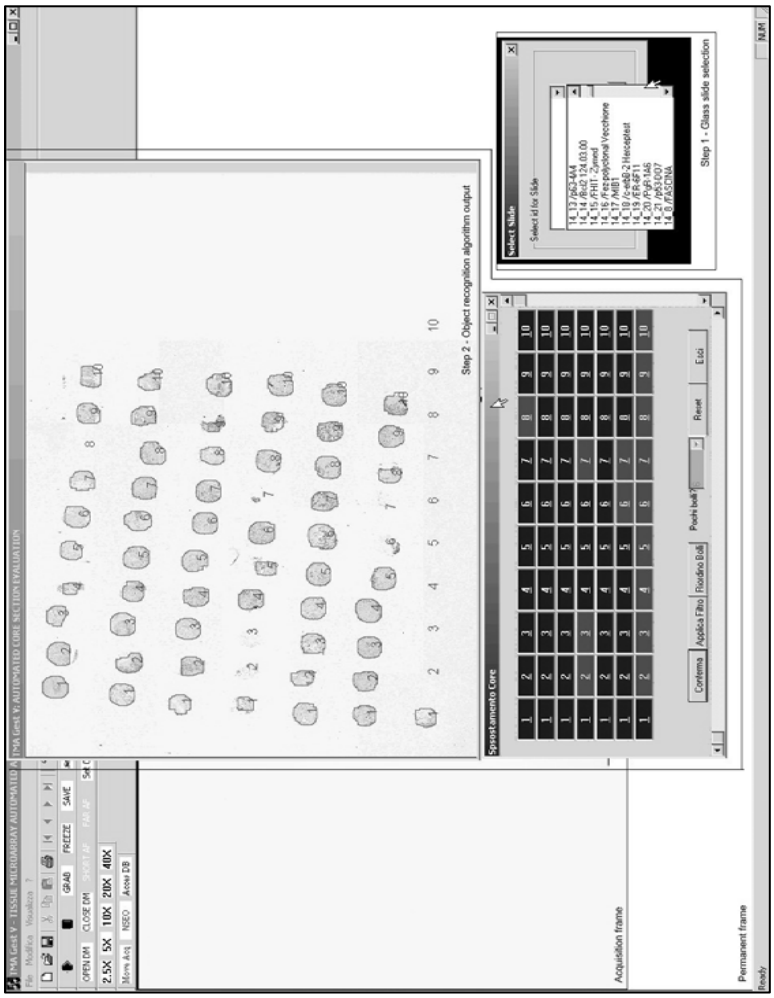


Figure 10 - Digital TMA Environment user interface: Step 1 frame corresponds to the already acquired slide selection to be processed by the object recognition algorithm; Step 2 frame represents the object recognition output. Through the lower frame the output may be corrected or the algorithm may be run again with different parameters.

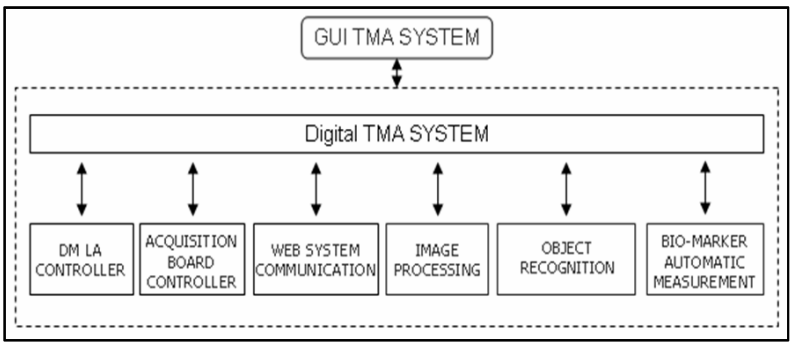


Figure 11 - Digital TMA Environment

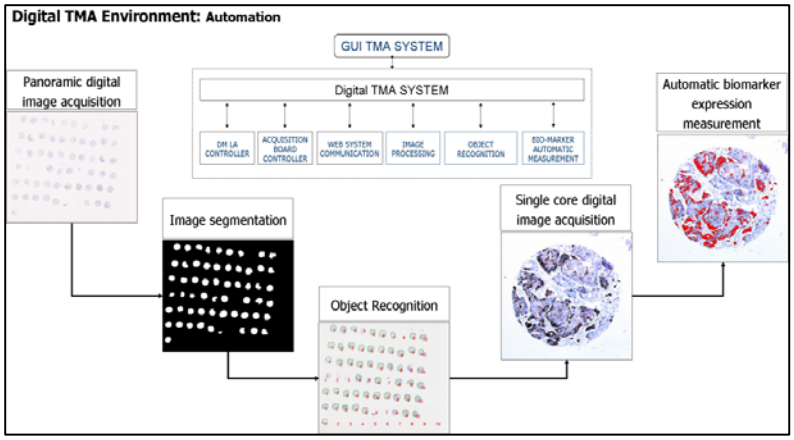


Figure 12 - TMA slide automatic acquisition: from panoramic overview detection to single spot automatic evaluation.

Panoramic acquisition and centre of coordinates

A crucial issue of correctly creating this image archive in an automated way is that all stored single core section images must be properly linked to the specimen (donor block) from which the core was punched and therefore to the underlying clinical-pathological information.

The entire automated process we propose as solution may be divided into four fundamental steps: (1) acquisition of a low-resolution digital image of entire glass slide (panoramic image); (2) identification of centre coordinates of each core section on panoramic image; (3) assignment of the correct position of each core in the matrix which topologically describes the microarray; (4) acquisition of single core section images at proper resolution in a correct ordered sequence.
Each step is performed automatically. User intervention in passing from step (3) to step (4) is optional. In next sections steps (1)-(4) are described in more detail. A screenshot of the Digital TMA Environment interface is shown in Figure 10.

Panoramic digital image acquisition

Low-resolution digital images of whole TMA slides are acquired capturing multiple regions of the slide. All digital images are 758x570pixel, 24 bit, acquired at 2.5 magnification (NA = 0.07).
These images are then digitally tiled together to create the resulting composite whole slide image (panoramic image). Before tiling, each image is automatically processed for shading correction. Image shading might be caused for example by non-uniform illumination or non-uniform camera sensitivity; its elimination is recommended when image analysis and understanding are final goals. With our equipment the resolution of panoramic images is about 35μm per pixel (approximately 50 images are tiled).

Identification of centre coordinates

A digital image processing procedure automatically identifies the position of the centre of each core section in the panoramic image. It works on the red component of colored panoramic image; thresholding and morphological operators are applied. A frame and an area filter are used together with a control on the number of segmented objects to automatically reduce the risk of detecting unwanted objects (noise) in the output binary image (for example

pen signs on a slide edge or air bubbles). These error sources are avoided firstly by using a rectangular frame to reject objects on the glass slide outside the region occupied by the array, and secondly by employing a bandpass area filter. The frame is determined during panoramic image acquisition step and surrounds the array region on the glass slide. The area filter automatically removes too small or too large objects from the array region, whenever the number of segmented objects exceeds the declared number of core section (reference value for core section area comparison is computed by $\pi(d/2)^2$, where d is the diameter of the needle). Once this procedure is completed, x and y coordinates of the centre of gravity of each core section are available (see Figure 13a-d), as output of image processing procedure.
The success of this feature extraction procedure is guaranteed if the values of two acquisition parameters, lamp intensity and diaphragm aperture, vary in predefined ranges of values.

Object recognition and ordering algorithm

The TMA construction procedure should produce regular block array sections, in which each core section position is easily recoverable from x- and y-spacing parameters. However, inconveniences occurring during block array sectioning and glass slide preparation complicate the object recognition procedure. Some examples of this kind of problems are reported in Figure 14. During the image processing procedure, some samples may be lost due to fragmentation, because of their too small dimensions, therefore contributing to the disorder of the panoramic digital image. Nevertheless, even these cases contain precious information which an automated procedure should not loose.
Therefore, efforts have been devoted to cover and solve these “disordered cases” in a way as more general as possible. The result of object recognition procedure is twofold: (1) the ordered list of the coordinates of the core section centres and (2) an index as-

sociated to each array site (grid cell), which summarizes additional information. This information allows the user to detect cases for which grid assignment is doubtful. Mis-assignments may induce propagation errors in the ordering procedure. The user may benefit of this information to manually correct the ordering results.
The object recognition procedure has a high degree of accuracy; therefore, in a large scale experiment as a TMA experiment is, we could use it with no manual intervention and automatically rejecting the few single core sections which are not properly located into the array. However, we developed an interactive module to recover these core sections.
The interactive module also allows one to change some of the parameters of the segmentation and the object recognition routines. The user can change the range of bandpass area filter to recover cases in which the automatic procedure was not able to remove all noise. Moreover, core section diameter on the glass slide is sometimes larger than the declared value, especially in the case of 1mm needle diameter. We found empirical reduced values for x- and y-spacing parameter which work better in case of enlarged core sections. The manual module allows switching between declared and empirical x- and y-spacing values. Whenever one of these parameters is changed, a new run of the segmentation and object recognition is executed.
The interactivity of our procedure is limited only to this manual module, which use is optional.

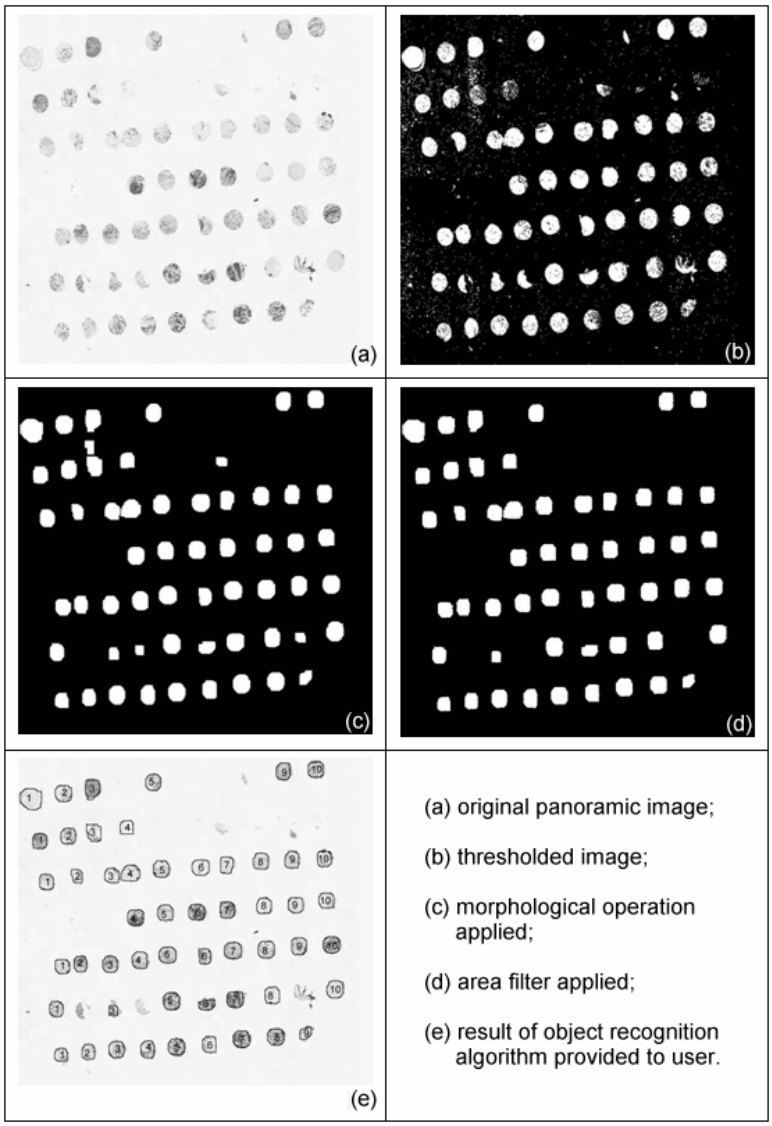


Figure 13 - Image processing procedure and result of object recognition algorithm on the panoramic image of a TMA arranged into a

single matrix of 7 rows and 10 columns.

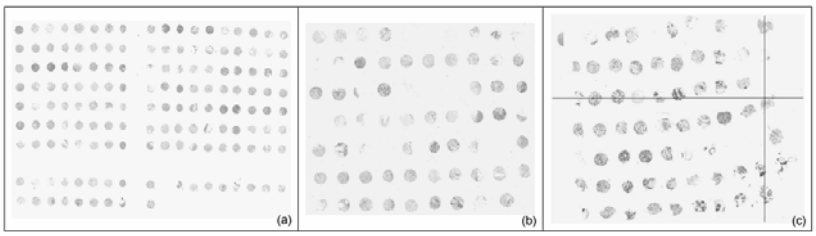


Figure 14 - Examples of panoramic images: (a) regular block array section (4 sub-matrix topology); (b) array with fragmented and lost core sections (single matrix topology); (c) distorted block array section: lines highlight distortions different from overall grid rotation.

Acquisition of single core sections image

The ordered sequence of core section coordinates is the input of the single image acquisition procedure. The stage of the robotic microscope is automatically driven on the glass slide and stopped on each core section to acquire the corresponding digital image at selected magnification. TMA digital cases are therefore constructed by allowing the user to examine the single core section at the required magnification. This approach is similar to digital cases acquisition, where tissue images are more straightforward acquirable because tissue is continuous. Magnification selection depends both on pathologists' needs (screen evaluation) and on image processing routine constraints, which automatically quantify biomarkers expression (quantitative image analysis).
Panoramic and single core section digital images are stored into the database.

Routine validation

To test the digital image processing and object recognition procedure 85 panoramic images of breast cancer TMA, created at St. Chiara Hospital, Trento, Italy were acquired at ITC-irst, Trento,

Italy. Seventy-five sections were cut from block arrays prepared with 1 mm diameter needle and arranged into a 7x10 single matrix (for an example, see Figure 14). The remaining sections originate from block arrays containing roughly 200 cores of 0.6mm diameter arranged in 4 sub-matrixes (for an example, see Figure 14). Selected sections were cut from a total of 29 different block arrays.
We identified the worst cases (17 cases) and used them as training set both for segmentation and object recognition algorithm. Concerning the segmentation we optimized digital acquisition parameters (lamp intensity, diaphragm aperture) on the training set and tested them on the remaining 68 images (test set). Concerning the procedure for the identification of centre coordinates, we report that the panoramic digital image acquisition takes about 4 minutes for about 50 tiles (array dimension about 20x22mm).
The segmentation algorithm correctly identified all core sections in all images. In fact, the optimal acquisition parameters found on the 17 training images allowed to successfully threshold all the test images in an automated way. For 12% of them the area filter was necessary to remove unwanted objects.
Hereinafter we discuss the results of object recognition algorithm performed with the 68 test images previously mentioned. The perfomance values are summarized in Table 1.

Table 1 - Object recognition algorithm performances.

	Core sections (%)
Number of grid cells	5878
Number of correct assignments	5692 (96.84)
Propagation errors	186 (3.16)
Localization errors	49 (0.83)

TMA automatic evaluation

Image analysis routines are integrated in the system, to fully automatically evaluate biomarker expression on single core sec-

tions. Automatic evaluations can then be stored in the database. Image processing algorithms are differently developed for specific cell compartment biomarker. These algorithms are not discussed in this work.
After a slide has been digitized, automated evaluation can be performed online (immediately after acquisition) or offline (later on), by choosing the slide code on a list of the already acquired slides and by choosing the automatic routine to work with. Each slide can be analyzed independently multiple times and all automatic outputs are stored in the database. Multiple evaluations can be necessary, for instance for markers staining both the cell membrane and the stroma, or to compare new implemented routines.
The image analysis routines included in the Digital TMA Environment do not require technicians to set calibration parameters, which could affect successive expression measurements.
However, the tissue heterogeneity (see section 5.3.1) may affect results of a fully automated evaluation and pathologist supervision is required to verify core section tissue.
Details of the image analysis routines are not provided herein. All the work regarding image processing analysis was done by R. Dell'Anna (ITC-irst, Trento, I).

4.2.3 System usage and availability

In 2002 a prototype version of the system was in use in the department of Pathology of Santa Chiara Hopital, Trento, Italy. Since June 2003 the web version of TMABoost is in use. As of January 2005, there are 7 institutions (from Italy, UK, USA) that have used the system to evaluate TMA experiments. About 2500 tumor cases (breast, ovary, and lung) are stored in the database and approximately 38000 TMA images have been evalauted. About 30 biomarkers have been evaluated on TMAs.
Current studies are devoted to the assessment of the prognostic power of novel biomarkers for breast and ovarian cancer. The prognostic significance of the Herceptest® on breast carcinomas on TMA was recently assessed [66].

A finite time period demo access to the web system can be requested at the web site [20]. About 20 demo accesses have been allowed till December 2004.

The object recognition and the ordering algorithm solution were presented at the CNIO Meeting 2003 [21] and a full paper was submitted to Computer Methods and Programs in Biomedicine. The TMABoost system was presented at Medinfo 2004 [60] and BITS 2004 [67] and a full paper was submitted to IEEE Transactions on Information Technology in Biomedicine.

Chapter 5

5. TMA Data Analysis

Recently a significant amount of work has been done in the area of data analysis for expression micorarrays. This work has focused on preprocessing approaches (for example, different flavors of normalization and filtering techniques) and on unsupervised and supervised methods based on statistical or machine learning techniques. Interestingly, many of these approaches are applicable to TMA studies, where panels of proteins are evaluated on cohorts of cases. All the critical issues of these methods were recently reviewed [68][44][69].

At a very high level, differences between expression microarray and TMA (see section 3.3) experiments can be mainly divided into qualitative and quantitative differences. The qualitative difference has to do with the kind of biological information involved in the experiments: contrarily to cDNA and oligonucleotide microarrays, immunohistochemical experiments with TMAs measure levels of proteins, and thus directly providing functional information of diseased cells *in situ*. As immunohistochemistry is routinely used in pathology laboratories throughout the world, TMA studies have the potential to be more translatable to a clinical application such as the development of a diagnostic biomarkers or a potential to therapeutic target. The quantitative differences have to do with number of variables involved in experiments and with quantity of data produced, which are much more controllable and manageable. Variables into play are more likely to be under control (antibody specificity and tissue conditions) allowing for higher data quality.

Typically TMA studies involve histopathological and clinical variables (covariates), as when used to construct predictive mod-

els. The protein covariates are studied for association with clinical parameters (tumor grade, tumor stage, survival, recurrence), and used along with clinical covariates in Cox regression models; see [70][71] and references therein. The very aim is to improve the predicting power of current prognostic models [40][72][73][74], using appropriate measures to compare the predicting power of them.
Another application regards the investigation of proteins to become powerful diagnostic biomarkers in clinical setting or of panel of proteins evaluated on same cohort of patients to define molecular profiles typical of disease progression (classification task). It either can be performed by using diagnoses as training labels or by using clinical outcomes (continuous training variable [17][18][75][76]).
Moreover studies are focused on the discovery of unknown diagnostic sub classes (discovery driven studies).
TMAs also play an important role in validation studies. There are now many examples of expression microarray and proteomic studies where TMAs are used to verify that the over or under expressed genes of interest are in fact differentially expressed *in situ* [77].

In this chapter we focus on practical aspects peculiar of TMA studies, regarding data quality assessment, data preprocessing and data analysis. In this context we consider automation respect to data quality.

5.1. Preprocessing

Integrated frameworks for the collection and management of data are of primary importance (see chapter 4), in particular if different types of data and many data sources are involved at distinct time points. Technological approaches can dramatically reduce the lack of homogeneity and errors in raw data, by applying consistency constraints, completeness controls, etc. Examples are the consistency control of birth and death dates, of tumor organ site and histological diagnosis, of compatibility between block array

description data (patient replicates, rows and columns number) and number of donor blocks used during preparation, etc..
Nevertheless, beyond the implementation of such technological solutions which can partially solve the problem, data preprocessing still remains a crucial phase of data analysis.
Data cleaning, data homogeneity (distribution analysis), outlier detection, missing data replacement (imputing), normalization, etc. must be almost always performed and standard statistical techniques are well suited for these purposes. As a next step a qualitative overview of the dataset is a good step.
One significant advantage in TMA data evaluation is being able to perform a visual inspection of the experimental data. Visual inspection can even help in detecting unexpected variations, which might be caused by errors or might be of biological interest. When multiple proteins are under investigation, conventional graphical representation such as distribution plots, box plots, etc, become too time-consuming.
An excellent example of good intuitive graphical tools are heat maps, which have been extensively used in expression microarray data analysis [103]. They represent quantitative data in a color based approach, assigning incremental color intensities to even spaced value intervals of expression levels. Many features and many samples can thus be visualized in one shot (see Figure 30).
It becomes more useful and interpretable if agglomerative hierarchical clustering is performed grouping samples based on proteins and/or vice versa. Hierarchical clustering output is graphically represented by dendrograms, which help in ‘ordering’ the heatmap [103].
The partitioning clustering technique is another example of clustering [78], that explores for natural groups of samples in the protein space or vice versa. K-mean and k-metoid approaches both require the number of clusters (k) to be sought as input parameter, and work to maximize the between cluster variability and minimize the within cluster variability until convergence. As far as the number of cases is manageable, comparison among single case assignments obtained with different k-values can be done, making

use of class labels. Figure 15 shows results of examples of a set of 54 cases (belonging to 4 diagnostic categories), which were clustered by a partitioning algorithm with three different k-values (k = 2, 3, 4).

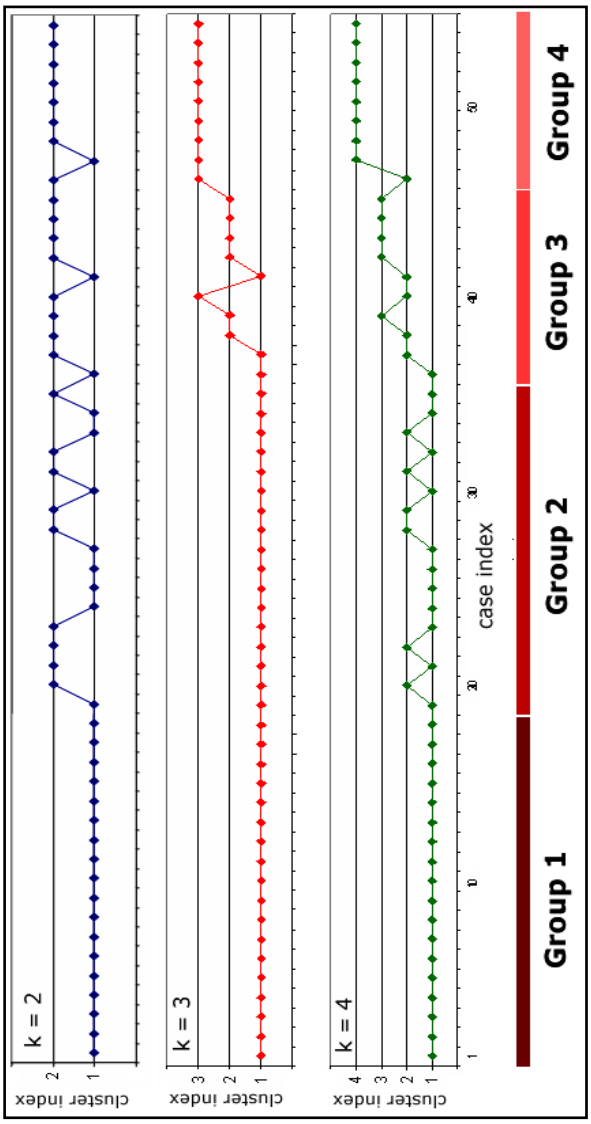


Figure 15 - Partitioning clustering: set of 54 cases (belonging to 4 diagnostic categories), clustered by a partitioning algorithm with three different k-values (k = 2, 3, 4).

5.2. Data reliability/reproducibility

Data reliability and reproducibility must be considered as part of the experimental design process and their assessment must be positioned before real experimental data acquisition. In the case of protein level detection by means of *in situ* antibody evaluation (on TMA samples), two approaches are adopted (see section 3.3.2.1), human or automatic evaluation. Assuming that sample preparation procedures are reproducible, inter- intra-observer variability must be assessed, both for human and for automatic evaluations.

5.2.1 Human evaluation

Assuming that evaluators/observers are equally skilled (equal expertise), inter-observer variability can be explained by subjectivity (it is difficult to have absolute reference for nominal categories) [79]; intra-observer variability is probably ascribable to the contingency of the situation [80].
In our experience, good agreements between pathologists in evaluating immuno-stainings are obtained only if data are dichotomized. We experienced it on two sets of more than 230 TMA core sections each, using kappa statistics (data are not provided).
It is reasonable to think that with a lot of training inter and intra-observer concordance could improve, but probably it will never be perfect.
If one of the final goals of studying proteins on tissues is to use biomarkers in clinical routine, inter-observer variability is going to be a challenging problem to be faced (see, for example [81]).

5.2.2 Automatic evaluation

Image processing routines are increasingly applied for the evaluation of gene and protein expressions, overcoming subjectivity problems and providing quantitative continuous data. Automatic

quantitative analysis algorithms are deterministic: given the same input, exactly the same output is expected.
By automatic evaluation intra-observer variability is thus cancelled and inter-observer variability (variability between two equipments) can be reduced or even avoided if accurate definition of settings is done.

Usually automatic machines work as 'black boxes' and, so far as no human intervention is required, data reproducibility is certain. If it is not the case and human interaction is required (semi-automatic systems), reproducibility must be assessed.
We experienced the use of a commercial semi-automatic system for the acquisition and evaluation of digital images and protein evaluations on TMA samples (color probe based immunohistochemistry). Quantitative data were provided by the image processing software, which evaluated four quantities for each core section (spot) despite the type of the marker (nuclear, membrane, etc.) (brown area detected, blue area detected, percentage of stained area, intensity). As the system has not means of distinguishing non tumor tissue from tumor tissue (benign, stroma, inflammation, etc), a pathologist reviewed all images. If needed, he circled the tumor tissue areas by hand and immediately reran the evaluation routine.

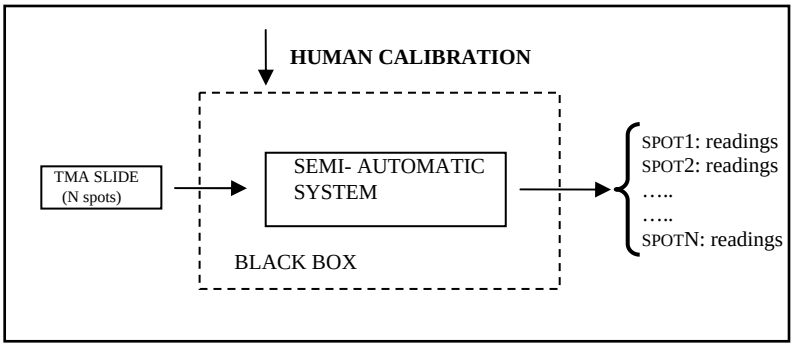


Figure 16 – Semi-automatic environment.

We investigated this point asking five pathologists to separately make calibrations, provided the circling was the same. We looked at the outputs and at their discrimination of groups. Data, reported in

A unique default setting is proposed by the company, but it is far from being suitable for each experiment.

For each experiment (each slide) calibration/setting must be provided by a pathologist at the beginning of the automated evaluation run (see Figure 16). The user is allowed to set many different parameters (Hue, Luminosity, and Saturation for Light Brown, Blue and Dark Brown) to calibrate the machine. This operation is based on his/her personal experience. Subjectivity comes again into play, being a potential drawback of this semi automatic system.

How do output data distributions and evaluation ranges change respect to initial settings? How do these changes affect further analyses (marker discriminative power against histotype-groups)?

We investigated this point asking five pathologists to separately make calibrations, provided the circling was the same. We looked at the outputs and at their discrimination of groups. Data, reported in

Appendix A, demonstrate that initial settings significantly affect discriminative power of biomarkers in distinguishing diagnosis groups.

This result warns to carefully handle semi-automatic quantitative systems to obtain reliable data.

5.3. TMA data peculiarities

In addition to data reliability assessment requirement, two major sources of irregolarities are present when dealing with raw TMA data:

We investigated this point asking five pathologists to separately make calibrations, provided the circling was the same. We looked at the outputs and at their discrimination of groups. Data, reported in - TMA Data Analysis

- tissue heterogeneity (topological). Not biologically interesting, but essential to be monitored (see the drilling problem, section 5.3.1).
- protein expression heterogeneity (proteins are generally not equally expressed across tumoral tissue). This heterogeneity might be of biological interest (see the pooling problem, section 5.3.2)

5.3.1 Drilling problem

Tissue heterogeneity can cause changes in diagnoses going down into TMA paraffin block depth; it turns out that different cuts might contain different information. It obviously can happen with conventional slides too, but it is less likely to be a problem, because larger areas are available for evaluations. Moreover, conventional slides are typically evaluated by pathologist, rather than automatically.

On the contrary, automation in TMA based studies is almost compulsory and is becoming more and more common.

How do these changes affect data analysis? We analyzed the ‘drilling problem’ in prostate tissues and statistically compared results obtained on a bunch of biomarkers with and without pathologist supervision of diagnoses (target diagnosis set versus pathologist diagnosis set), obtaining statistically different results.

It suggests that for spatially heterogeneous tissues, pathologist supervision to confirm hysto type of single spot is mandatory and that, therefore, the experimental procedure can not be totally automatic. A solution to this problem might be provided by alternative quantitative automatic approach as [57] so far as appropriate biomarkers to select tissue compartments are available [22].

Figure 17 shows how diagnoses change along the punches (z-axis of block arrays) in a prostate TMA.

We investigated this point asking five pathologists to separately make calibrations, provided the circling was the same. We looked at the ou[illegible] Data, reported in

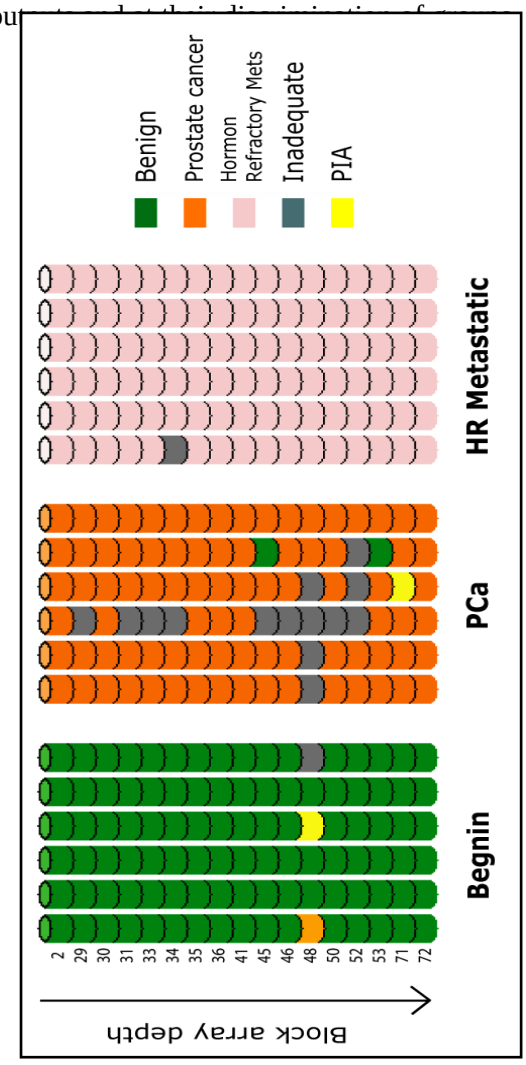


Figure 17 – Loss of target tissue along 72 TMA experiments: the figure represents one row along z-axis.

We investigated this point asking five pathologists to separately make calibrations, provided the circling was the same. We looked at the outputs and at their discrimination of groups. Data, reported in - TMA Data Analysis

5.3.2 Pooling problem

Usually multiple punches from the same donor block (patient) are included in a block array both to improve the possibility to have valuable samples of each patient (even if some construction problems occurred, see section 3.3.1) and to better represent the biological variability of the protein expression.

Several studies were proposed on this topic, trying to quantify the most appropriate number of replicates for different tumor types [27].

When TMA cuts are then stained multiple spots of each patient are evaluated for each protein. Figure 18 schematically represents how multiple markers TMA dataset looks like: data entries along one column do not contain information at the same level.

Each data entry is of the form y_{ab}^{l}, expression level of protein *b* (marker) on patient *a* (case), read on replicate *l*. Number of replicates can vary from case to case and marker to marker.

How do these replicates be managed in further analysis? No universal rule exists. Replicates are usually pooled [40]: minimum, maximum, or mean value among replicates is used. Standard deviations from the mean value for each case can be evaluated as measure of spread of the protein expression.

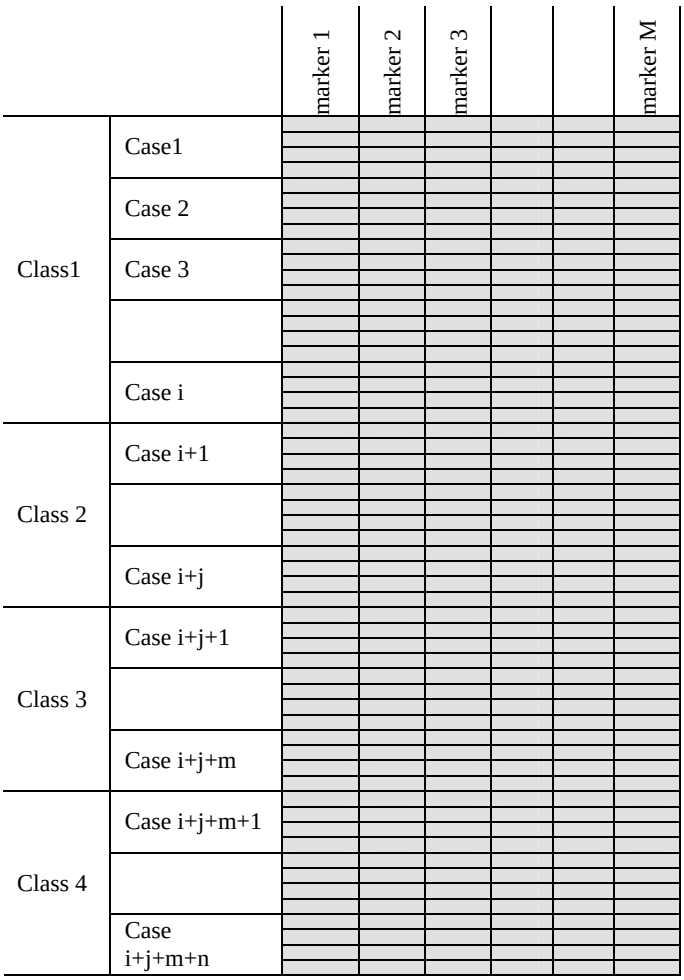
Pooling approach might be decided on the basis of prior biological knowledge (pooling method should reflect knowledge about the biological action of the protein). For example, for the estrogen receptors ER marker, which expression is known to significantly improve patient outcome, ER positivity in one spot provides necessary information, thus the maximum value among replicates can be chosen.

But it is not often the case and alternative approaches should be applied.

A reasonable alternative relies on the associations of the investigated biomarker with one or more other parameters (biological, histological or clinical) being relevant for the study domain. Dif-

We investigated this point asking five pathologists to separately make calibrations, provided the circling was the same. We looked at the outputs and at their discrimination of groups. Data, reported in

ferent pooling choices may lead to different results, as we (data not provided) and others [104] experienced, at least as regards the statistical significance of the association. This approach strongly requires a validation step on other datasets and if the ‘choice’ holds, it must be preserved in the future (a kind of protocol).



We investigated this point asking five pathologists to separately make calibrations, provided the circling was the same. We looked at the outputs and at their discrimination of groups. Data, reported in - TMA Data Analysis

Figure 18 –TMA dataset structure.

To account for multiple measures of same case and marker, we propose a Bayesian hierarchical classification approach, described in chapter 6.

5.4. Expression value dichotomization

It is quite common [40] to dichotomize expression scores of biomarker evaluation, when interested in characterize low risk and high risk patients (both in univariate and multivariate analysis). The rationale is twofold: i. leading with pathologists' evaluations, inter-observer variability is definitely high (dichotomization overcomes this problem) and ii. over or under expression of a biomarker is of easier biological interpretation.

The drawback is that if you deal with automatic evaluated measures, which are quantitative and continuous, dichotomization wastes a lot of information that might be of interest. It is usually the case that anyhow, not enough cases are available to use continuous values in survival analysis. Provided that the continuous spectrum of values is of interest, a compromise would be the definition of ranges of values.

In any case when dealing with dichotomization (or even discretization), appropriate cutoff values should be defined. Similar considerations as for replicate pooling choices regarding biological prior knowledge (see 5.3.2) are here applicable [40]. Alternatively to Cox regression models, tree-predictors techniques can be applied [82] [83], leading to intuitive interpretable rule based models.

We investigated this point asking five pathologists to separately make calibrations, provided the circling was the same. We looked at the outputs and at their discrimination of groups. Data, reported in

Chapter 6

6. Bayesian Hierarchical Model

6.1. Introduction

Usually multiple punches from the same donor block (patient) are included in a block array both to improve the possibility to have valuable samples of each patient and to better represent the biological variability of the protein expression. Depending on tissue type and on the protein under study, intra-tumor variability is lower or comparable to inter-tumor variability (class variability).
Variations among replicate expression values are commonly ignored in TMA data analysis by straightforward pooling approaches. The mean, the maximum or the minimum among replicates is usually adopted and the strategy is based on biological knowledge or on protein associations (see section 5.3.2).
However it has been found [40], that different choices can lead to covariates with different significance levels in Cox regression [70].
Intra-tumor expression heterogeneity can provide additional interesting information to the analysis task. In a probabilistic framework for example, accounting for intra-tumor variability, could alter the posterior probability of a case of belonging to a certain class (or could even change the predicted class), providing insight into the particular case study.

When measurement occurs at different levels (see Figure 18), standard statistical techniques are not appropriate, because they either assume that groups belong to entirely different populations or ignore the aggregate information entirely.
Hierarchical models provide a way of pooling the information for the different groups without assuming that they belong to precisely the same population [43]. Hierarchical models (also called

multilevel models) are used when information is available at different levels of observation units (as for example, in meta-analysis of separate randomized trials).
In our application field, TMA studies, each group represents a patient or a tumor. Elements of the group are the multiple measurements of a protein evaluated on different spots (replicates). Multiple groups belong to the same class, for instance, a diagnostic class. A schematic view of the data structure is represented in Figure 19.

Herein we propose a classification model, which accounts for intra-tumor variability in a probabilistic framework. It is based on a one-way normal random effects model that is a special case of the hierarchical normal linear model.
We applied a Bayesian approach for dealing with inter-measurement variability. Such an approach is perfectly suited for managing, within the standard decision theory, the uncertainty affecting the experimental data.
We compare the performances of our model $\mathbf{M}_{\text{HierBa}}$ to a standard Bayesian classifier model $\mathbf{M}_{\text{StBa}}$. The $\mathbf{M}_{\text{StBa}}$ model is a Bayesian model for which we applied standard pooling strategies. Performances are evaluated on simulated datasets characterized by different ratios of intra-variability over inter-variability (within and between groups variability). Performance evaluations are obtained for a one feature two class problem. We evaluated our approach also on a real dataset.
Some considerations on the proposed approach are drawn at the end of the chapter.

The model was implemented in the R statistical package [107].

6.2. Model definition

Be x_{ij} the value of the replicate j of case i, where $j = 1,....,n_{rep}$ and $i = 1,....,N_{C_k}$, k being the class index and

N_{C_k} being the number of cases. Assume that for each case i, the x_{ij} are normally distributed, $x_{ij} \sim N(\mu_i, \sigma_i^2)$ and that replicate variances are all the same, $\sigma_i^2 = \sigma^2$ for each i. Also assume that the replicate means of each case μ_i are normally distributed, $\mu_i \sim N(M, \tau^2)$. The linear hierarchical model is sketched in Figure 19.
In the next sections we firstly describe the classification task and secondly the learning task.

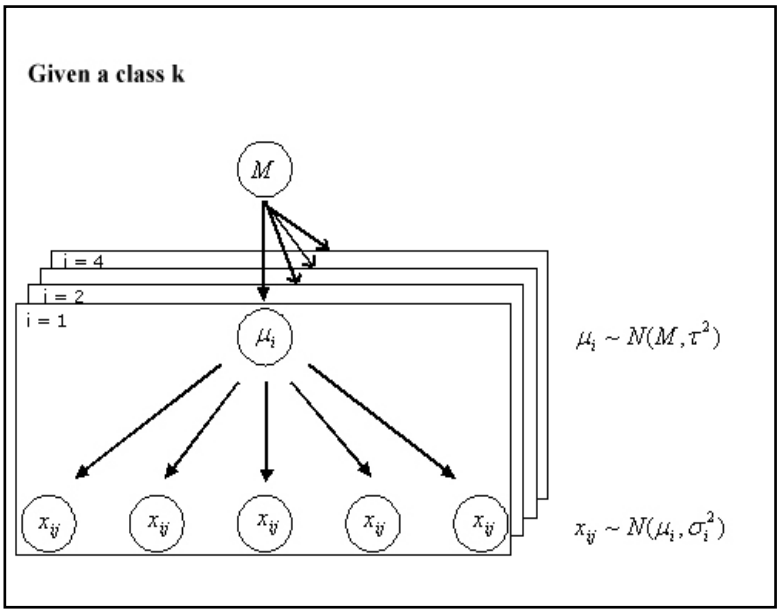


Figure 19 - Structure of the hierarchical model.

6.2.1 M_{HierBa} Model : Classification

By Bayes's theorem, the posterior probability for a new case, $\mathbf{X} \in \mathrm{R}^{n_{rep}}$, is

$$P(C_k \mid \mathbf{X}) = \frac{P(C_k)P(\mathbf{X} \mid C_k)}{P(\mathbf{X})} = \frac{P(C_k)P(\mathbf{X} \mid C_k)}{\sum_k P(C_k)P(\mathbf{X} \mid C_k)}.$$

To evaluate the posterior probability, the marginal likelihood must be solved:

$$P(\mathbf{X} \mid C) = \int_{\mu} P(\mathbf{X} \mid \mu, \sigma^2) P(\mu \mid M, \tau^2) d\mu,$$

the k-index of the class is omitted for simplicity.

Under these conditions (normal distributions and equal replicate variances),

$$P(\mathbf{X}|C) = \frac{1}{\left(\sqrt{2\pi}/S\right)^{n_{rep}} \left(\sqrt{2\pi}/T\right)} \int_{\mu} \exp\left(-\frac{S^2}{2} \sum_j (x_j - \mu)^2 - \frac{T^2}{2} (\mu - M)^2 \right) d\mu,$$

where $S^2 = 1/\sigma^2$ and $T^2 = 1/\tau^2$.

Applying simple algebra,

$$P(\mathbf{X}|C) = \frac{1}{\left(\sqrt{2\pi}/S\right)^{n_{rep}} \left(\sqrt{2\pi}/T\right)} \int_{\mu} \exp\left(-\frac{S^2}{2} (\sum_j x_j^2 + n_{rep}\mu^2 - 2\mu \sum_j x_j) - \frac{T^2}{2} (\mu^2 + M^2 - 2\mu M) \right) d\mu$$

$$= \frac{\exp(-\frac{1}{2}(S^2\sum_j x_j^2 - T^2M^2))}{\left(\sqrt{2\pi}/S\right)^{n_{rep}}\left(\sqrt{2\pi}/T\right)} \int_\mu \exp\left(-\frac{1}{2}(S^2 n_{rep}\mu^2 - 2S^2\mu\sum_j x_j + T^2\mu^2 - 2T^2\mu M)\right)d\mu$$

$$= \frac{\exp(-\frac{1}{2}(S^2\sum_j x_j^2 - T^2M^2))}{\left(\sqrt{2\pi}/S\right)^{n_{rep}}\left(\sqrt{2\pi}/T\right)} \int_\mu \exp\left(-\frac{1}{2}(S^2 n_{rep} + T^2)(\mu^2 - 2\mu\frac{(S^2\sum_j x_j + T^2\mu)}{(S^2 n_{rep} + T^2)})\right)d\mu$$

$$= \frac{\exp(-\frac{1}{2}(S^2\sum_j x_j^2 - T^2M^2))}{\left(\sqrt{2\pi}/S\right)^{n_{rep}}\left(\sqrt{2\pi}/T\right)} \exp(\frac{(S^2 n_{rep}\, x_{mean} + T^2 M)^2}{2(S^2 n_{rep} + T^2)})$$

$$\int_\mu \exp\left(-\frac{1}{2}(S^2 n_{rep} + T^2)(\mu - \frac{(S^2 n_{rep}\, x_{mean} + T^2\mu)}{(S^2 n_{rep} + T^2)})^2\right)d\mu$$

$$= \frac{\exp(-\frac{1}{2}(S^2\sum_j x_j^2 - T^2M^2))}{\left(\sqrt{2\pi}/S\right)^{n_{rep}}\left(\sqrt{2\pi}/T\right)} \exp(\frac{(S^2 n_{rep} x_{mean} + T^2 M)^2}{2(S^2 n_{rep} + T^2)}) \frac{\sqrt{2\pi}}{\sqrt{S^2 n_{rep} + T^2}},$$

we obtain

$$P(\mathbf{X}|C) = \frac{\sigma}{\left(\sqrt{2\pi}\sigma\right)^{n_{rep}}\sqrt{n_{rep}\tau^2 + \sigma^2}} \exp(-\frac{\sum_j x_j^2}{2\sigma^2} + \frac{M^2}{2\tau^2}) \exp(\frac{(\frac{\tau^2 n_{rep}^2 x_{mean}^2}{\sigma^2} + \frac{\sigma^2 M^2}{\tau^2} + 2n_{rep} x_{mean} M)}{2(n_{rep}\tau^2 + \sigma^2)})$$

A new instance will be classified into the class with maximal posterior probability.

Generalization for the multidimensional case is easily obtainable, assuming the conditional independence of the features given the class.

6.2.2 M_{HierBa} Model : Learning

We estimated the distribution parameters from the data. This approach is usually called empirical learning [84][85]; it is an approximation of the maximum likelihood estimation.
Within groups, means and variances are straightforward evaluated as:

$$\hat{\mu}_i = \frac{\sum_j^{n_{rep}} x_{ij}}{n_{rep}} \text{ and } \hat{\sigma}_i^2 = \frac{\sum_j (x_{ij} - \mu_i)^2}{n_{rep} - 1}.$$

Between groups we applied the pooled estimate for the class mean:

$$\hat{M}_{pool} = \frac{\sum_i^{N_{C_k}} \frac{\mu_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}.$$

Between groups variance is estimated as

$$\hat{\tau}_{corr}^2 = \frac{\sum_i (\mu_i - \hat{M}_{pool})^2}{N_{C_k} - 1} - \frac{\sum_i \sum_j (x_{ij} - \mu_i)^2}{N_{C_k}(n_{rep} - 1)}.$$

The variance $\hat{\tau}_{corr}^2$ includes two terms, respectively for *between* groups and *within* groups variances (inter-variability and intra-variability, respectively). It is like purifying the variance by the measure variability.

This point estimate of the variance is valid as far as the within group variability is comparable to the between group variability, i.e. as far as the population variance is comparable to single case variance.

These choices (complete pooling) are reasonable when the ratio of between to within mean squares is not significantly greater than 1 (see [43]).

6.2.3 M_{StBa} Model : Classification

For a standard Bayesian classifier, the likelihood of x, $x \in \mathrm{R}$, is

$$P(x \mid C) = \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{1}{2\tau^2}(x - M)^2\right),$$

being τ and M, the variance and the mean of the population distribution.

By Bayes's theorem, we evaluate the posterior probability for a new instance, $x \in \mathrm{R}$. The value x of each instance corresponds to the mean value among replicates, x_{mean} (standard pooling strategy).
A new instance will be classified into the class with maximal posterior probability.

6.2.4 M_{StBa} Model: Learning

We estimated distribution parameters (means and variances) as $\hat{M} = \frac{\sum_{i}^{N_{C_k}} \hat{\mu}_i}{N_{C_k}}$ and the variance $\hat{\tau}^2 = \frac{\sum_{i} (\hat{\mu}_i - \hat{M})^2}{N_{C_k} - 1}$.

6.3. Synthetic data

Simulated datasets had 2000 patients (1000 each class) and 5 replicates each patient.
All expression values were generated as normal random numbers with fixed variances (σ_k^2) and with means being normal random numbers with means 90 and 140 and fixed variances τ_k^2 .

6.4. Validation of the M_{HierBa} model

To validate the model we performed three sets of experiments with different aims.

I. CLASSIFICATION TASK: It is aimed to verify the performances of the classifier M_{HierBa}, under different distribution conditions for the two classes. We assumed model parameters as known.
Experiments were made on synthetic data, with:
$\tau_{k=1}^2 = \tau_{k=2}^2$ and $\sigma_{k=1}^2 = \sigma_{k=2}^2$ (balanced);
$\tau_{k=1}^2 \leq \tau_{k=2}^2$ and $\sigma_{k=1}^2 \leq \sigma_{k=2}^2$ (unbalanced), with different values of τ_k^2 and σ_k^2.

II. LEARNING TASK: It is voted to verify the empirical learning reliability (point estimation of the parameters). Classification performances are compared to the previous. Experiments were made on synthetic data.

III. LEARNING: As previously, but on a real set of data.

For all the experiments, the priors for the classes were set at 0.5 and the decision threshold at 0.5, unless differently specified. Results of the experiments are reported below.

I set of experiments

Ia.

Evaluation of classification performances of the two models $\mathbf{M}_{\text{HierBa}}$ and $\mathbf{M}_{\text{StBa}}$, equally varying the variances for the two classes for a bunch of parameter sets ($\tau^2_{k=1} = \tau^2_{k=2}$ and $\sigma^2_{k=1} = \sigma^2_{k=2}$). Table 2 contains the data of the 8 runs we performed. It reports the generating variances for the two classes, the confusion matrix of the classification, the accuracies and the mean value of the absolute differences of the posteriors (for each case). With balanced data no differences respect to the classification accuracy are expected (by classification model definition).

Table 2 – Classification performances on balanced simulated datasets. No learning applied.

	Class 1, MCl1 = 90		Class2, MCl1 = 140		$\mathbf{M}_{\text{HierBa}}$ and $\mathbf{M}_{\text{StBa}}$			Mean of abs diff posteriors
	$\tau^2_{k=1}$	$\sigma^2_{k=1}$	$\tau^2_{k=2}$	$\sigma^2_{k=2}$	Conf matrix: [target, model]		Accuracy	
1	600	100	600	100	848	152	0.849	0.0043
					150	850		
2	600	200	600	200	821	179	0.827	0.0086
					166	834		
3	600	300	600	300	837	163	0.838	0.0129
					160	840		
4	600	400	600	400	847	153	0.844	0.0172
					159	841		
5	600	500	600	500	834	166	0.845	0.0211
					143	857		
6	600	600	600	600	819	181	0.821	0.0247
					176	824		
7	300	100	300	100	926	74	0.923	0.0053
					80	920		
8	300	300	300	300	907	93	0.913	0.0167
					81	919		

Ib.

Evaluation of classification performances of the two models $\mathbf{M}_{\text{HierBa}}$ and $\mathbf{M}_{\text{StBa}}$, unequally varying the variances for the two

classes for a bunch of parameter sets ($\tau^2_{k=1} \leq \tau^2_{k=2}$ and $\sigma^2_{k=1} \leq \sigma^2_{k=2}$). Table 3 contains the data of the 8 runs we performed. It reports the generating variances for the two classes, the confusion matrix of the classification, the accuracies and the mean value of the absolute differences of the posteriors (for each case). Figure 20 shows the posterior probabilities of the cases (panel (a)) and the values of the misclassified cases by the two models (panel (b)). Data of experiment 5 (Table 3) are plotted.

Table 3 - Classification performances on unbalanced simulated datasets. No learning applied.

Exp	Class 1, MCl1 = 90		Class2, MCl1 = 140		$\mathbf{M}_{\text{HierBa}}$		$\mathbf{M}_{\text{StBa}}$		$\mathbf{M}_{\text{HierBa}}$	$\mathbf{M}_{\text{StBa}}$	Mean of abs diff posteriors
	$\tau^2_{k=1}$	$\sigma^2_{k=1}$	$\tau^2_{k=2}$	$\sigma^2_{k=2}$	Conf matrix [target, model]		Conf matrix [target, model]		Acc.	Acc.	
1	300	100	600	100	911	89	911	89	0.895	0.895	0.0054
					121	879	121	879			
2	300	100	600	200	931	69	909	91	0.892	0.863	0.0638
					146	854	183	817			
3	300	100	600	300	939	61	909	91	0.920	0.878	0.0885
					98	902	152	848			
4	300	100	600	400	953	47	902	98	0.938	0.872	0.1172
					76	924	157	843			
5	300	100	600	400	951	49	913	87	0.933	0.879	0.1104
					84	916	154	846			
6	300	100	600	500	958	42	912	88	0.949	0.877	0.1235
					59	941	158	842			
7	300	100	600	600	956	44	901	99	0.946	0.859	0.1439
					64	936	183	817			
8	600	100	600	400	918	82	826	174	0.911	0.814	0.1520
					95	905	197	803			

Table 4 - Classification performances on unbalanced simulated datasets. Parameters were learned from data.

Exp.	Class 1, MCl1 = 90		Class2, MCl1 = 140		Conf matrix: [target, model]		Conf matrix: [target, model]		Accuracy		Mean of abs diff posteriors
	$\tau^2_{k=1}$	$\sigma^2_{k=1}$	$\tau^2_{k=2}$	$\sigma^2_{k=2}$	$\mathbf{M}_{\text{HierBa}}$		$\mathbf{M}_{\text{StBa}}$		$\mathbf{M}_{\text{HierBa}}$	$\mathbf{M}_{\text{StBa}}$	
1	300	100	600	400	949 83	51 917	914 156	86 844	0.933	0.879	0.1095
2	600	100	600	400	904 103	96 897	821 194	179 806	0.900	0.813	0.1567

II set of experiments

We performed the empirical learning of the distribution parameters. Experiments are performed on the datasets of exp.5 and exp.8 of Table 3, two situations of unbalanced distributed data. Data are reported in Table 4.

The estimated parameters in experiment 1 are:

$\hat{M}_{k=1}$ = 89.78; $\hat{M}_{pool,k=1}$ = 89.00; $\hat{\tau}^2_{corr,k=1}$ =221.25; $\hat{\sigma}^2_{mean,k=1}$ = 97.60;

$\hat{M}_{k=2}$ =140.07; $\hat{M}_{pool,k=2}$ =140.36; $\hat{\tau}^2_{corr,k=2}$ =266.59; $\hat{\sigma}^2_{mean,k=2}$ =398.44.

The estimated parameters in experiment 2 are:

$\hat{M}_{k=1}$ =90.66; $\hat{M}_{pool,k=1}$ = 89.32; $\hat{\tau}^2_{corr,k=1}$ =534.54; $\hat{\sigma}^2_{mean,k=1}$ = 99.67;

$\hat{M}_{k=2}$ =138.14; $\hat{M}_{pool,k=2}$ =137.15; $\hat{\tau}^2_{corr,k=2}$ =281.02; $\hat{\sigma}^2_{mean,k=2}$ =410.72.

We applied cross-validation with different choices of fold number. Accuracy values are reported in Table 5.

Table 5 – Classification performances for unbalanced datasets evaluated by cross validation.

Exp	Fold number	Accuracy (StDev) $\mathbf{M}_{\text{HierBa}}$	Accuracy (StDev) $\mathbf{M}_{\text{StBa}}$
As 1	10	0.932 (0.018)	0.88 (0.018)
As 1	5	0.932 (0.012)	0.88 (0.016)
As 2	10	0.90 (0.02)	0.85 (0.018)
As 2	5	0.904 (0.012)	0.85 (0.009)

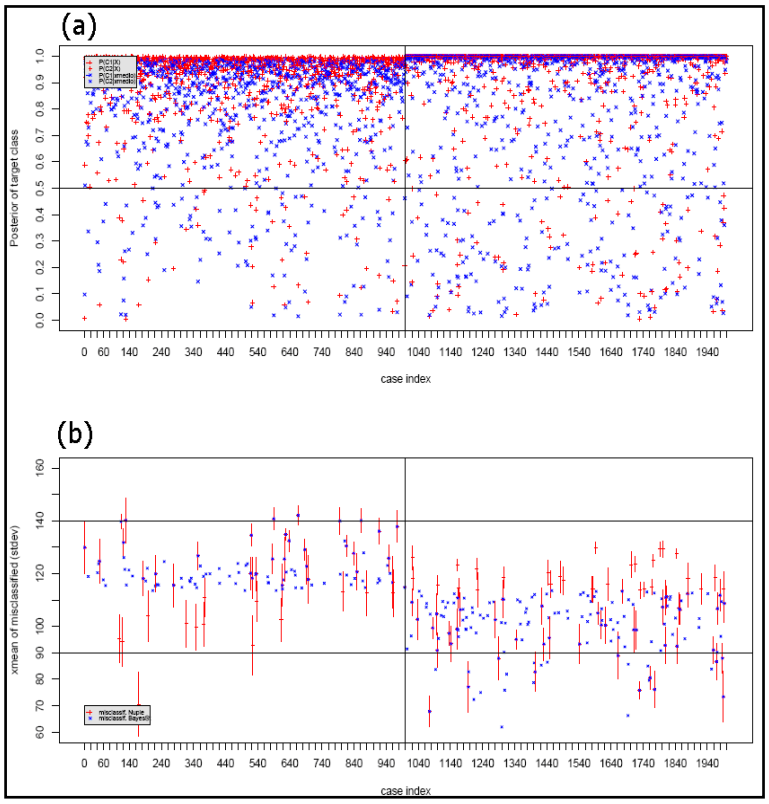


Figure 20 - Posterior probabilities (panel (a)) and misclassified data (panel (b)) of the models $\mathbf{M_{HierBa\ and}}$ $\mathbf{M_{StBa}}$ are plotted respect to the case index. No learning applied. Cases with index *i* ∈ [1, 1000] have target class C_1 and with index *i* ∈ [1001, 2000] have target class C_2. Panel (a) shows posterior probabilities of the cases, i.e. $P(C_k \mid x_{mean,i})$ for $\mathbf{M_{StBa}}$ and $P(C_k \mid \mathbf{X}_i)$ for $\mathbf{M_{HierBa}}$. Panel (b) shows misclassified data by each of the two models: x_{mean} values for the $\mathbf{M_{StBa}}$ are reported (blue) and x_{mean} values and standard deviations for the $\mathbf{M_{HierBa}}$ (red). Horizontal lines correspond to the mean values of the two classes.

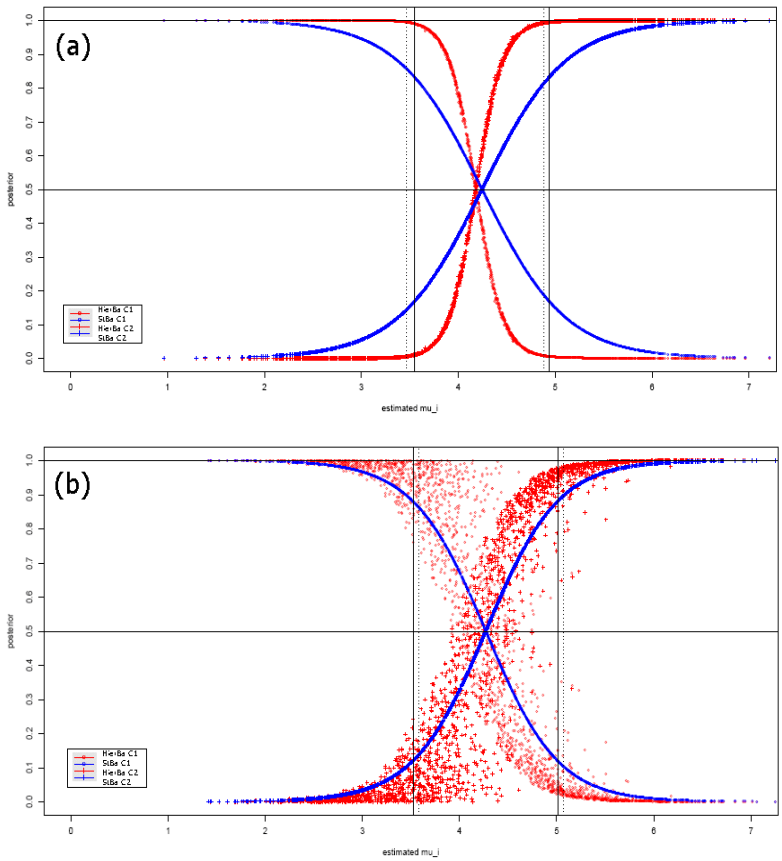


Figure 21 - Posteriors are plotted against the evaluated means of the cases (x_{mean}). Priors were evaluated from data. Panel (a) shows data of a balanced dataset ($\hat{\tau}^2_{k=1,2} = 0.5, \hat{\sigma}^2_{k=1,2} = 0.5$). Panel (b) shows data of an unbalanced dataset ($\hat{\tau}^2_{k=1,2} = 0.5, \hat{\sigma}^2_{k=1} = 0.2, \hat{\sigma}^2_{k=1} = 0.4$). Vertical lines identify the values of the estimated class means $\hat{M}$ by the learning approaches of M_{HierBa} and M_{StBa}.

III set of experiments

We applied our classification model to a prostate dataset composed of 122 benign prostate tissues and 205 localized prostate cancer tissues. The α-Methylacyl CoA racemase (AMACR) protein was evaluated on this dataset.

AMACR is a sensitive recently identified prostate cancer biomarker [22]. It is consistently over expressed in prostate cancer both at transcriptomic level and at proteomic level. AMACR is lower expressed in metastatic prostate cancer compared to localized prostate cancer, as recently confirmed by fluorescent-based measurements [22].

The distributions of the AMACR dataset are shown in Figure 22. For the two classes, the case replicate data, the case variances and the means are reported. The estimated parameters of the two classes are:

$\hat{M}_{k=1}$=3.39; $\hat{M}_{pool,k=1}$=3.15; $\hat{\tau}^2_{k=1}$=0.48; $\hat{\tau}^2_{corr,k=1}$=0.135; $\hat{\sigma}^2_{mean,k=1}$=0.41;

$\hat{M}_{k=2}$=4.35; $\hat{M}_{pool,k=2}$=4.59; $\hat{\tau}^2_{k=2}$=0.58; $\hat{\tau}^2_{corr,k=2}$=0.27; $\hat{\sigma}^2_{mean,k=2}$=0.36.

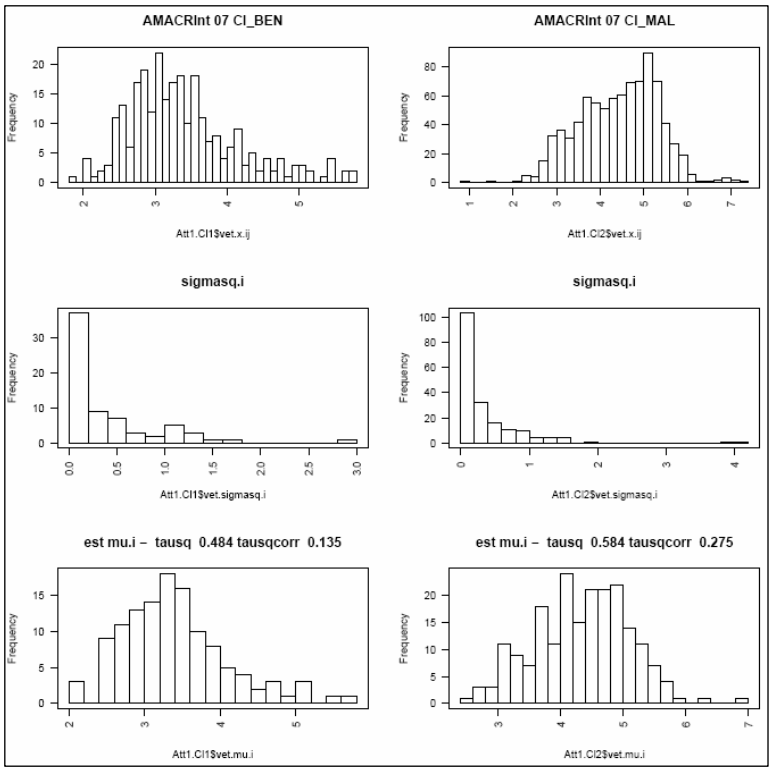


Figure 22 - The α-Methylacyl CoA racemase AMACR dataset distributions. Core section evaluations (row 1), variance values $\hat{\sigma}^2$ (row 2) and means among replicates $\hat{\mu}$ (row 3) are plotted.

Table 6 shows the performances on the whole dataset and Table 7 shows the results obtained applying cross validation.

Table 6 - Classification performances on AMACR dataset. Exp. 2 performed with priors evaluated from data (0.37, 0.63 respectively); exp 3 with decision threshold for C_1 equal to 0.7 and equal priors.

Exp.	Conf matrix: [target, model]		Conf matrix: [target, model]		Accuracy		Mean of abs diff posteriors
	$\mathbf{M}_{HierBa}$		$\mathbf{M}_{StBa}$		$\mathbf{M}_{HierBa}$	$\mathbf{M}_{StBa}$	
1	99	23	100	22	0.7584	0.7523	0.1155
	56	149	59	146			
2	91	31	86	36	0.7676	0.7859	0.1113
	45	160	34	171			
3	87	35	70	52	0.7737	0.7584	0.1155
	39	166	27	178			

Table 7 - Classification performances on AMACR dataset by applying cross validation.

Exp.	Folds	Accuracy (StDev) $\mathbf{M}_{HierBa}$	Accuracy (StDev) $\mathbf{M}_{StBa}$
1	2	0.77 (0.04)	0.77 (0.06)
2	5	0.79 (0.07)	0.78 (0.08)
3	10	0.73 (0.10)	0.72 (0.11)

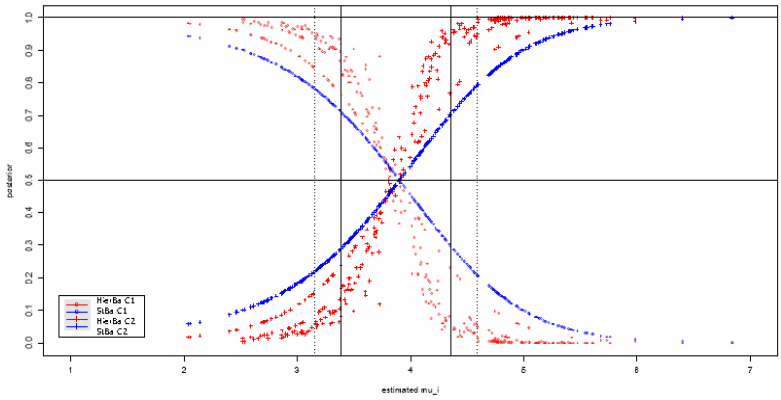


Figure 23 – AMACR dataset posteriors plotted against the evaluated mean of the cases (x_{mean}).

6.5. Considerations

These preliminary performance tests of the $\mathbf{M}_{\text{HierBa}}$ give us some clues about the applicability of the model.
The following considerations apply to a one feature two class problem.
Regarding the classification model evaluation we observed that when classes are balanced (see I set of experiments) no differences in terms of classification rates are obtained (intrinsic into the model); however increasing differences in the posteriors are detected as the intra-variability (within group) increases $P(C_k \mid \mathbf{X}) \neq P(C_k \mid xmedio)$, for each *k*.
As far as the classes are unbalanced the two models $\mathbf{M}_{\text{HierBa}}$ and $\mathbf{M}_{\text{StBa}}$ perform differently (see Table 3). As the difference in the class variances becomes larger, the $\mathbf{M}_{\text{HierBa}}$ outperforms the $\mathbf{M}_{\text{StBa}}$. Misclassifications of the two models are different, as shown in the panel (b) of Figure 20.
We then tested the reliability of the empirical learning on two unbalanced datasets (see II set of experiments). We observed that the models have lower but still good performances (compare Table 3 and Table 4) and that they decrease more for the $\mathbf{M}_{\text{HierBa}}$ than for the $\mathbf{M}_{\text{StBa.}}$. This can be explained by the fact that the hierarchical model includes more parameters than the standard model, incrementing the learning effort. The observations made for the classification model evaluation still hold.
When dealing with the real dataset (III set of experiments), we observe that classification performances of the two models are almost the same. Differences are present in terms of single case posterior as shown in Figure 23.

Based on these very preliminary tests we can conclude that the proposed model works at worse as the standard Bayesian classification model and that under some constraints it works better. The empirical learning of the parameters provides reasonable results.
The interesting part of the approach is related to the posterior probability evaluations. Our model emphasizes the information

available at every level, accounting for the spread of the replicate measures. An extreme example is that our model differentiates between vectors as $\mathbf{X}_1 = (90,90,90,90,90)$ and $\mathbf{X}_2 = (90,90)$ (awarding $\mathbf{X}_1$), whereas the standard approach does not.

Future work will face the evaluation of the performances on other real datasets with different characteristics. Simultaneously the model will be test in a multi-features framework. Successively a complete Bayesian approach can be implemented.

The proposed model is applicable on other domains.

Chapter 7

7. Protein Expression in Human Cancer

Here we report results from three datasets examining different biological or clinical questions. For each study we briefly describe the motivation of the study, the dataset, the analysis technique applied and the results obtained. The first two studies were aimed to investigate the role of two proteins, M-CAM and Jagged1, as prognostic markers for ovarian and prostate cancer, respectively. The third study evaluated the possibility of building a model to discriminate among different states of prostate cancer starting from a panel of 41 proteins; this third study was built on data from previous findings obtained using a combination of proteomics and expression array analysis. Unsupervised approaches were applied to qualitatively investigate the data. Supervised approaches were adopted to build a classification model.

7.1. M-CAM expression in ovarian carcinomas

This study was conducted in collaboration with the Department of Experimental Oncology, National Cancer Institute, Milano, I, the Department of Histopathology, S. Raffaele Hospital, Milano,I, and the Department of Histopathology, Gynecological Surgery and Medical Oncology, S. Chiara Hospital, Trento, I.

M-CAM expression in ovarian carcinomas is histologically restricted to serous and undifferentiated tumors and is an independent marker of poor prognosis.

Ovarian cancer is the most common gynecological malignancy causing fatality in western countries and the fifth leading cause of female cancer death. Despite progress in surgical and chemotherapy treatment, the 5-year survival rate of all stages of ovarian

cancer has remained constant at 39% over the past 30 years. Traditional clinico-pathological criteria used to predict clinical outcome are inadequate. Thus, great benefit is likely to result from the characterization of additional prognostic factors, more closed related to tumor biology. These biological factors may offer novel approaches to the identification of groups of patients that could benefit from more aggressive therapy or could become targets for a more rationale therapy. M-CAM is classified as a member of the Ig-CAM superfamily on the basis of the homology of the nucleotide sequence [86] Cell-adhesion molecules (CAM) are involved in tissue morphogenesis and in tumor progression of several human neoplasms. Although the biological role of M-CAM in normal tissue and malignant tumors remains unclear, M-CAM has been suggested to play an important role in tumor progression, implantation and placentation [86].
The aim of the present study is to investigate the expression of M-CAM in a large series of well characterized human samples of ovarian carcinomas, and to investigate if its expression at the tissue level is related to pathologic and clinic characteristics, prognosis and response to therapy. To do this we used a properly designed tissue microarray and investigated M-CAM expression using a specific monoclonal antibody.

7.1.1 Material and method

Patients and tumor specimens. To analyze the clinical and pathological significance of M-CAM expression we selected a series of formalin-fixed paraffin-embedded tissue blocks of 130 surgically resected ovarian carcinomas. All cases had been operated between 1992 and 1999. All histological sections and paraffin blocks were obtained from the Departments of Pathology of the S. Chiara Hospital of Trento and of the Borgo Trento Hospital of Verona. One pathologist with peculiar skill in gynecological pathology reviewed all pathological data. For all cases complete clinical data, including long term follow-up, were available from the Departments of Gynecological Surgery and Medical Oncol-

ogy of the S. Chiara Hospital of Trento and of the Borgo Trento Hospital of Verona. The series included serous papillary (82 cases), endometrioid (12), clear cell (13), indifferentiated (13), mucinous (6) and other (2), NA 5. Al cases have been staged according to FIGO criteria: 26 were stage I, 14 stage II, 68 stage 3 and 21 stage 4, and 4 NA. All cases were treated with platinum based therapeutic schedules. Response to therapy was known for 99 patients and was scored as complete (58 cases), partial (22), minimal (1) or absent (18). Clinico-pathological data of all patients and of the subgroups of serous papillary cases and of cases which showed complete or partial response to treatment are summarized in Table 8, Table 9 and Table 10 respectively.

Tissue microarray construction. One representative tissue block for each tumor was selected and 4 tumor areas were identified on the corresponding histological slide and dotted on the slide. These dots were used to generate TMA blocks using a manual TMA instrument (Beecher Instruments, Inc., Sun Prarie, WI). Each TMA was designed to include 4 punches (needle diameter 0.6 mm) from each paraffin block, as well as a series of appropriate controls.

Immunohistochemistry and scoring system. Five micron sections of the TMA were prepared and immediately processed. Prior to immunostainings, sections underwent heat induced epitope retrieval using citrate buffer, as previously described [87]. Immunostaining was performed using a specific anti M-CAM monoclonal antibody and the D07 anti p53 monoclonal antibody, followed by a sensitive StreptABC detection system as previously described [88]. Positive controls for M-CAM staining were the endothelial cells within the samples [89]. Positive controls for p53 were done on cases of breast carcinoma with known p53 overexpression [90]. Negative controls were obtained by omitting primary antibodies.

The results of the immunostainings were read by two different observers at the multiheaded microscope. For each core biopsy the observers evaluated the visual intensity of staining (0=no staining, 1=faint staining, 2=moderate staining, 3= intense stain-

ing) and the estimated percentage of reacting cells, which was subsequently categorized as follows: 0=non reacting cells, 1=1 to 10% of reacting cells, 2=11 to 25% of reacting cells, 3=26 to 50% of reacting cells, 4= more than 50% of reacting cells. Moreover we recorded the precise cellular localization of the reaction product, i.e: at the cell membrane or in the cytoplasm. A final immunohistochemical score for membrane or cytoplasmic staining was obtained by adding the value of the intensity to the percentage category. We also evaluated if there was immunostainings of the stromal component of the tumors, which was graded from 0 to 3.
All data concerning the tumor characteristics and the block array maps were previously stored in the TMAboost system (see chapter 4) so as the immunostainings results as described above (intensity and percentage for membrane, cytoplasma and intensity for stromal component). Immunohistochemical data were subsequently retrieved and analyzed in relation to pathological and clinical parameters and with relapse free (RFS) and overall survival (OS).
We analyzed separately the immunostainings of the four core sections of each case, and recorded each value separately: thus the immunohistochemical analyses generated a maximum of 24 different values for each case (for each of the 4 cores we recorded intensity, percentage and score for membrane and cytoplasmic staining and evaluated also stromal staining). This prompted us to investigate as to how classify each case in the over or under expressed protein group. Because no previous hints were available we aggregated the replicates values by using minimum, mean and maximum and looked for correlations with pathological and clinical parameters.

Statistical analysis. For statistical analyses beside studying the whole population, we focused on some more homogeneous subsets of patients which were grouped together on the basis of similar clinico-pathological parameters such as stage I and II, stage 3 and 4, complete and partial responders and minimally and no responders.

Chi-squared test was used to analyze the correlation of M-CAM positive cases according to clinicopathological characteristics. Fisher test was performed between dichotomous variables (two by two contingency tables).
Kaplan-Meier curves were used to estimate recurrence-free time and overall survival, and the log rank test served to assess whether curves differed significantly between groups. We performed the analyses on the whole series of tumors, and on subgroups of tumors with homogeneous serous papillary histology, stage and type of response to therapy.
Multivariate analysis. To analyze the recurrence risk and patient survival accounting for multiple clinicopathologic parameters and M-CAM expression, multivariate logistic regression was used. The Cox proportional hazards model was used to assess which covariates affected recurrence-free time and overall survival.
For all analyses P-values less than or equal to 0.05 were considered statistically significant. Analyses were performed with the software package R [107].

Table 8- Pathological and clinical characteristics of the whole tumor series of ovarian neoplasms and their relation to M-CAM expression.

		Total	M-CAM negative	M-CAM positive	P value
Mean age (57.75)			57.25	59.85	ns
	<mean	57	34	23	
	>=mean	68	31	37	ns
Stage	Stage I and II	40	31	9	
	Stage III	68	24	44	
	Stage IV	21	11	10	< 0.001
Histotype	Serous papillary	82	35	47	
	Indifferentiated	13	3	10	
	Clear cell	13	12	1	
	Endometrioid	12	9	3	
	Mucinous	6	5	1	
	Other	2	2	0	< 0.001
Grading	Grade 1 and 2	61	36	25	
	Grade 3 and 4	63	27	36	ns
Response to	Responders	80	37	43	

therapy					
	Non responders	19	12	7	ns
P53	Score 0-4	62	42	20	
	Score 5-7	68	23	45	< 0.001

Table 9 - Pathological and clinical characteristics of the serous papillary ovarian carcinomas and their relation to M-CAM expression.

		Total	CD146 negative	CD146 positive	P value
Mean age (59)	<mean	31	13	18	
	>=mean	47	21	26	ns
Stage	Stage I and II	15	10	5	
	Stage III	51	17	34	
	Stage IV	14	7	7	0.058
Grading	Grade 1 and 2	41	19	22	
	Grade 3 and 4	40	16	24	Ns
Response to therapy	Responders	42	20	32	
	Non responders	14	9	5	Ns
P53	Score 0-4	34	19	15	
	Score 5-7	47	15	32	0.053

Table 10 - Pathological and clinical characteristics of partial/complete responders tumors and their relation to M-CAM expression.

		Total	CD146 negative	CD146 positive	p value
Mean age (57.75)	<mean	38	20	18	
	>=mean	39	16	23	ns
Stage	Stage I and II	21	18	3	
	Stage III	49	13	36	
	Stage IV	10	6	4	< 0.001
Histotype	Serous papillary	52	20	32	
	Indifferentiated	6	2	4	
	Clear cell	6	5	1	
	Endometrioid	8	5	3	
	Mucinous	3	2	1	
	Other	2	2	0	0.1240
Grading	Grade 1 and 2	37	21	16	
	Grade 3 and 4	39	14	25	0.1111
P53	Score 0-4	39	24	15	
	Score 5-7	40	12	28	< 0.01

Table 11 – M-CAM, univariate survival analysis.

Variable	Grouping	P value	HR	95%CI
OS in the whole series of cases				
Stage	I and II vs III and IV	<0.00001	7.84	3.6-17.1
Age	<median;>=median	0.024	1.65	1.06-2.57
Response to therapy	Absent and min vs others	<0.00001	0.165	0.0944-0.288
Histotype	Serous vs all	0.00383	2.12	1.26-3.56
Grade	I and II vs III and IV	0.00494	1.91	1.21-3.02
P53	MaxScore07 04vs57	0.104	1.44	0.924-2.25
M-CAM	MaxCat 0vs123	0.000276	2.3	1.45-3.65
RFS in the whole series of cases				
Stage	I and II vs III and IV	<0.00001	7.88	3.9-15.9
Age	<median;>=median	0.412	1.19	0.775-1.84
Response to therapy	Absent and min vs others	<0.00001	0.179	0.102-0.316
Histotype	Serous vs all	0.00398	2.1	1.25-3.52
Grade	I and II vs III and IV	0.00451	1.91	1.21-3.02
P53	MaxScore07 04vs57	0.0147	1.74	1.11-2.72
M-CAM	MaxCat 0vs123	0.00099	2.12	1.34-3.35
OS in the group of responders				
Stage	I and II vs III and IV	<0.00001	15.4	3.75-63.7
Age	<median;>=median	0.123	1.54	0.883-2.68
Histotype	Serous vs all	0.002	2.99	1.44-6.21
Grade	I and II vs III and IV	0.0283	1.89	1.06-3.38
P53	MaxScore07 04vs57	0.155	1.50	0.854-2.62
M-CAM	MaxCat 0vs123	<0.00001	4.71	2.46-9.01
RFS in the group of responders				
Stage	I and II vs III and IV	<0.00001	18.5	4.48-76.8
Age	<median;>=median	0.628	1.14	0.657-1.98
Histotype	Serous vs all	0.00173	3.03	1.46- 6.28
Grade	I and II vs III and IV	0.103	1.61	0.904-2.87
P53	MaxScore07 04vs57	0.0728	1.68	0.947-2.97
M-CAM	MaxCat 0vs123	<0.00001	4.48	2.32-8.65

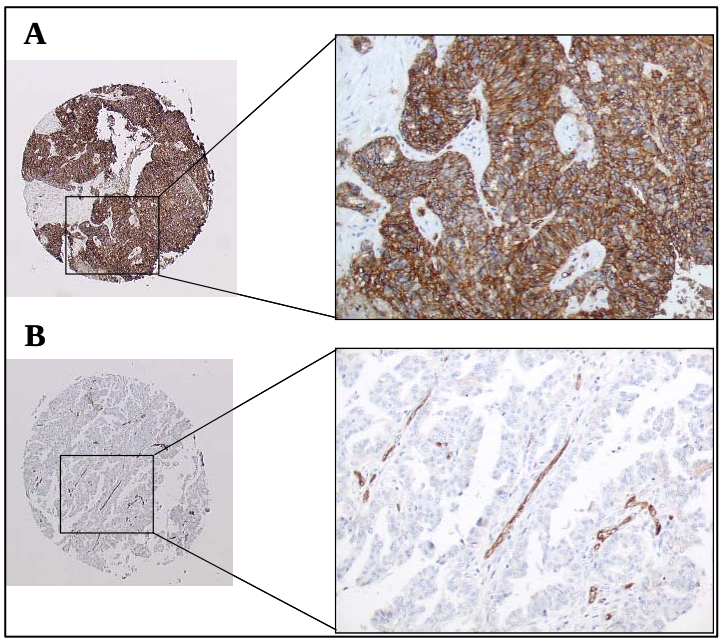


Figure 24 – Representative immunostainings of M-CAM in ovarian tissues.

7.1.2 Results

Patient population. The distribution of histological type, tumor stage and grade of the cases studied in this report roughly reflects the demographics of Italian population. Median follow-up was 48 months. Response to therapy was available in 98 cases: 78 (79.6%) cases were complete or partial responders, while 20 (20.4%) were minimally or non responders. High stage (I and II versus III and IV), high grade (1 and 2 versus 3 and 4), histological type (serous vs. other histotypes), and response to therapy were all associated with shorted relapse free and overall survival (RFS, OS) (
Table 11). Age was associate with worse OS only in the whole series of cases. High p53 expression (p53 score higher than 5)

was associated with shorter RFS in the whole series of cases and in the subset of serous tumors (
Table 11).

M-CAM expression. M-CAM was expressed in neoplastic cells in 88 (66.1%) of the investigated tumors. Expression was mainly seen at the cell membrane (65 cases, 48.9%) and less frequently in the cytoplasm (27 cases, 20.3%). Membrane and cytoplasmic expression were strictly associated (P<0.00001). M-CAM was always expressed by endothelial cells and in 29 cases it was also expressed in stromal non vascular tissue. In M-CAM positive tumors the percentage of cells with membrane reactivity ranged from 5% to 90% (median 30%) and staining intensity was distributed as follows: intensity 1 in 23 (17.3%) cases, 2 in 37 (27.8%) cases, and 3 in 5 (3.7%) cases. The M-CAM immunohistochemical score, showed following distribution: score 0 in 68 (51.1%) cases, 1 in 0 (0%) cases, 2 in 4 (3%) cases, 3 in 17 (12.8%) cases, 4 in 9 (6.8%) cases, 5 in 12 (9%) cases, 6 in 18 (13.5%) cases, and 7 in 5 (3.8%) cases.

M-CAM expression and clinico-pathological parameters. For statistical analysis, after extensive work to define the most appropriate and bias-free criterion (data not shown), tumors were considered positive for M-CAM if they had a score of 1 or higher in at least one of the TMA cores. M-CAM positivity was almost restricted to serous papillary and undifferentiated carcinomas (P=0.0004, see Tab. I), and it was associated with higher stage (P=0.0001, Table 8) and with high p53 levels (P=0.0002, Tab I).When only serous tumors were analyzed, M-CAM positivity retained only borderline statistically significant association with higher stage and p53 over-expression (P=0.059 and P=0.053,
Table 9). When only complete/partial responders were analyzed, M-CAM positivity was statistically significant association with higher stage and p53 over-expression (Table 10).

M-CAM expression and patient survival. At univariate survival analysis, positive M-CAM was associated with shorter RFS and OS in the whole series of cases, as well as in the subgroup of serous carcinomas (Table 10, Figure 25). By analyzing the group of

78 patients classified as responder to therapy, M-CAM positivity was strongly associated with shortened RFS and OS (P<0.0001) in the subgroup of responders tumors (
Table 11, Figure 26).

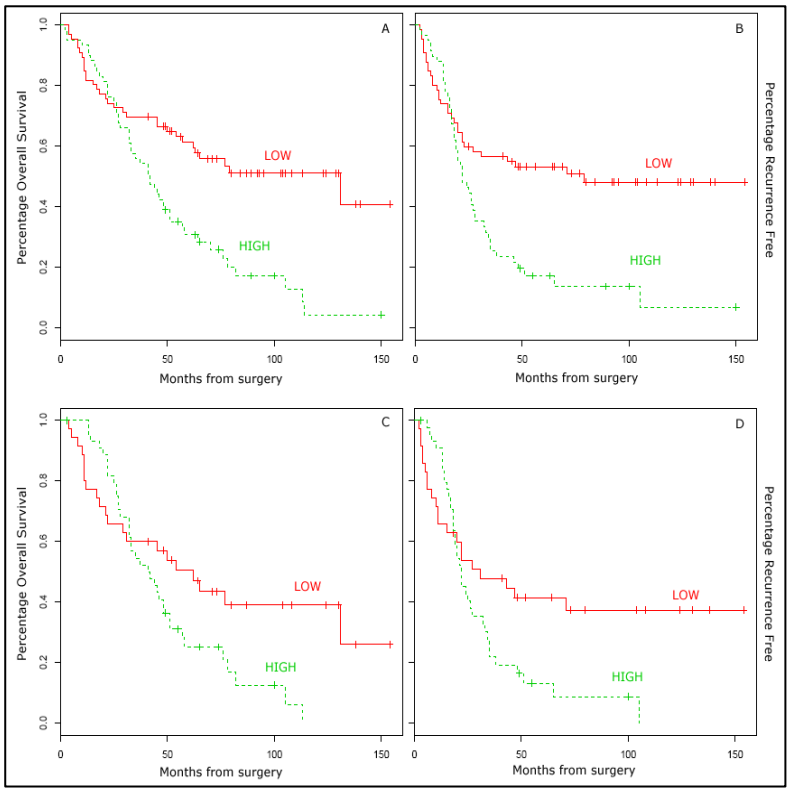


Figure 25 - M-CAM expression: A. and B. overall survival analysis and recurrence free analysis in the whole series, and C. and D. in serous adenocarcinomas.

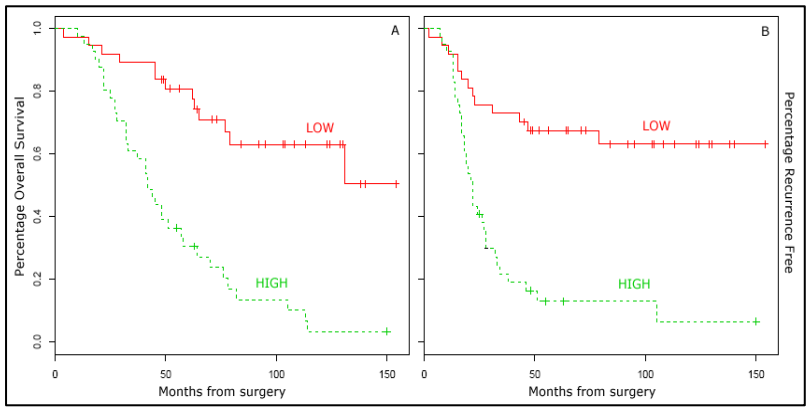


Figure 26 - M-CAM expression: A. and B. overall survival analysis and recurrence free analysis in responders.

Multivariate survival analysis. At multivariate analysis, high stage and response to therapy which were associated with both shortened RFS and OS in the whole series of patients and in the subgroup of serous carcinomas; age was and independent prognostic marker only for OS (Table 12). Analyzing the group of responder patients, high stage and M-CAM expression were independent prognostic markers predicting early relapse and shorter OS (Table 12). In a separate multivariate analysis not taking in account the grade of the lesions, M-CAM expression was independently associated with shorter RFS (P=0.0066).

OS in the whole series of cases				
Variable	Grouping	P value	HR	95%CI
Stage	I and II vs III and IV	<0.002	10.735	2.4046-47.927
Histotype	Serous vs all	0.88	1.056	0.5200-2.144
Grade	I and II vs III and IV	0.73	1.106	0.6303-1.941
Response to therapy	Absent and min vs others	<0.00001	0.191	0.0958-0.380
Age	<median;>=median	0.022	1.890	1.0945-3.262
M-CAM	MaxCat 0vs123	0.5	1.213	0.6925-2.124
RFS in the whole series of cases				
Stage	I and II vs III and IV	0.00079	13.164	2.924-59.260
Histotype	Serous vs all	0.60000	1.214	0.585-2.518
Grade	I and II vs III and IV	0.89000	0.961	0.533-1.734

Response to therapy	Absent and min vs others	0.00057	0.322	0.169-0.613
P53	MaxScore07 04vs57	0.24000	1.468	0.775-2.782
M-CAM	MaxCat 0vs123	0.57000	1.192	0.649-2.190
OS in the group of responders				
Stage	I and II vs III and IV	0.0270	5.68	1.219-26.45
Histotype	Serous vs all	0.1400	1.87	0.812-4.28
Grade	I and II vs III and IV	0.3500	1.34	0.724-2.48
M-CAM	MaxCat 0vs123	0.0081	2.65	1.289-5.44
RFS in the group of responders				
Stage	I and II vs III and IV	0.0025	9.95	2.245-44.10
Histotype	Serous vs all	0.2700	1.55	0.706-3.41
M-CAM	MaxCat 0vs123	0.0130	2.42	1.205-4.87

Table 12 - Multivariate survival analysis.

In the present study we demonstrate that M-CAM is expressed at the mRNA and protein levels in a relevant percentage of ovarian carcinomas. M-CAM is differentially expressed in the different tumor types, being associated with serous and undifferentiated histology. At the protein level, M-CAM expression is associated with higher stage of the tumors, and with poor prognosis. At multivariate analysis, we could also show that in patients, which respond to therapy, M-CAM expression independently identifies a group of patients at higher risk of early relapse and death.
Present data have multiple implications in ovarian cancer cell biology, histology and possibly selection of therapeutic targets. Based upon the known role of Ig-CAMs in cell adhesion, signaling, and tumor progression in other human neoplasm, it is tempting to hypothesize a role for M-CAM in ovarian carcinogenesis and progression. However to date we are not aware of experimental data on cell lines or animal model, which could support this suggestion, but our data suggest that further studies could be important in elucidating this point.

7.2. Jagged1

This study was conducted in collaboration with the Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, and the Children's Hospital Informatics Program, Children's Hospital Boston, Boston, MA; with the Dana Farber Cancer Institute, Boston, MA, and the Departments of Pathology and Urology, University of Michigan School of Medicine, Ann Arbor, MI.

Jagged1 Expression is Associated with Prostate Cancer Metastasis and Recurrence.
JAGGED1 is a NOTCH receptor ligand, recently identified [91] using a high throughput quantitative proteomic technique. This study describes the increased expression of Jagged1 protein in human prostate tumours and its associations with prostate cancer progression and metastasis. The findings support a model in which dysregulation of JAGGED1 protein levels plays a role in prostate cancer progression and metastasis and suggest that JAGGED1 may be a useful marker in distinguishing indolent and aggressive prostate cancers. Herein all what concerns the characterization of JAGGED1 antibodies (specificity test, see section 3.3.2.1) and its expression in cancers different than prostate is not reported. These aspects can be found in [92] together with more detailed biological discussion. In the following sections, analyses based on TMA experiments are reported.

7.2.1 Material and method

Case Selection
Samples from 236 patients with benign, high-grade prostatic intraepithelial neoplasia and localized and metastatic prostate cancer were obtained from the rapid autopsy program and the radical prostatectomy series, respectively, of the University of Michigan Prostate Cancer SPORE Tissue Bank and assembled on tissue mi-

croarrays as described previously (93, 94). A subset of 95 men with clinically localized prostate cancer was evaluated for associations between JAGGED1 expression levels and tumor recurrence (as judged by a rise in prostate-specific antigen level). A total of 1,247 tissue microarray cores were evaluated.

Immunohistochemical Analysis

Immunohistochemistry was done on 5-μm sections prepared from paraffin-embedded tissue microarrays. Slides were successively soaked in xylene; passed through graded alcohols; washed in distilled water; pretreated with 10 mmol/L citrate (pH 6.0; Zymed) in a steam pressure cooker (Decloaking Chamber, BioCare Medical, Walnut Creek, CA); and washed again in distilled water. All additional steps were done at 25°C in a hydrated chamber. Slides were pretreated with Peroxidase Block (DAKO USA) for 5 minutes, blocked with 20% goat serum for 20 minutes, and treated with rabbit anti-JAGGED1 antibody (H-114, Santa Cruz Biotechnology) at 1:50 in DAKO diluent for 1 hour. After washing in 50 mmol/L Tris-Cl (pH 7.4), antirabbit horseradish peroxidase-conjugated antibody (Envision detection kit, DAKO) was applied for 30 minutes. After additional washing, immunoperoxidase staining was developed with a DAB chromogen kit (DAKO). Slides were counterstained with hematoxylin.

Semiautomated Quantitative Biomarkers analysis

Evaluation of JAGGED1 immunohistochemical staining was done with the Chromavision (Chromavision Medical Systems, Inc., San Juan Capistrano, CA) Automated Cellular Imaging System (ACIS II). The system combines automated microscopy with computerized image analysis to generate a continuous immunohistochemical staining score between 0 to 255 intensity units. Each tissue microarray core was reviewed to ensure that intensity measurements were taken from diagnostic regions (benign, high-grade prostatic intraepithelial neoplasia, or cancer) and that staining scores obtained from non representative regions were excluded. Images of all tissue microarray cores used in this analysis,

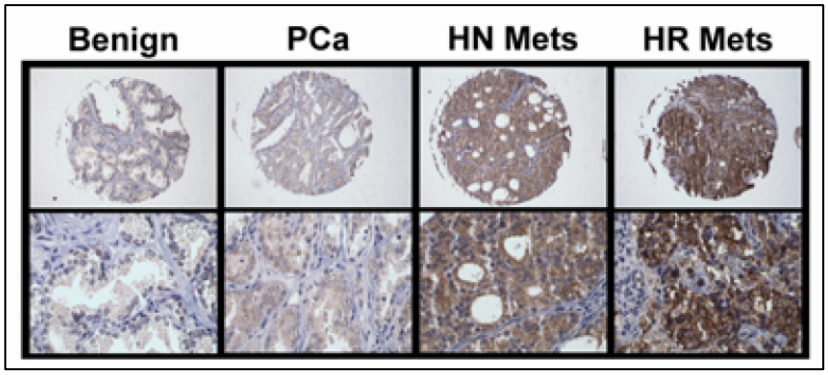
including those stained for JAGGED1, can be viewed at a supplementary web site (http://rubinlab.tch.harvard.edu/supplemental_data/JAGGED1/index.jsp).

Statistical Analysis**.**
Patient information, including pretreatment prostate specific antigen values, clinical and tumor stage, radical prostatectomy Gleason score, and surgical margin status, was prospectively collected and stored. After radical prostatectomy, patients were assessed annually for prostate specific antigen recurrence-free survival with a cut point of _0.2 ng/ml to define biochemical evidence of micrometastatic recurrence or progression. Bivariate (univariate) analysis was done to examine the association of clinical and pathological parameters with JAGGED1 staining and recurrence-free survival. Cox proportional hazards regression models were used to analyze the relationship between recurrence-free survival and preoperative variables. A backward selection procedure was implemented to select the most parsimonious model. A 0.05 significance level was used for all decisions of significance. Analyses were run with SPSS 11.0.1 (SPSS, Inc., Chicago, IL). Kaplan-Meier analysis was used to establish prostate-specific antigen recurrence-free survival.

7.2.2 Results

JAGGED1 Protein Expression Is Associated with Prostate Cancer Progression.
To study the expression of JAGGED1 in human prostatic specimens, immunostaining was done on high-density tissue microarrays containing benign prostatic tissues, localized untreated prostatic adenocarcinomas, and hormone-naive and -refractory metastatic tumors. Two samples from 18 different patients imprinted in duplicate were represented in these tissue microarrays (giving a total of 4 measurements per patient and 72 per category). The staining intensity of JAGGED1 was measured and

quantified with the Chromavision Automated Cellular Imaging System II. For each patient the minimum, maximum, and mean staining intensity among replicates of the same patient were



evaluated.

Figure 27 - Representative JAGGED1 immunostaining in prostatic tissue samples.

Representative examples of staining for JAGGED1 in benign prostatic tissue, localized cancer, and metastatic tumor are shown in Figure 27. Mean JAGGED1 staining intensity was increased significantly in clinically localized prostate cancer (score = 94.2; SE = 1.8) *versus* benign prostate tissue (score = 79.6; SE = 2.8; $P < 0.001$) and again in metastatic tumor (score = 127.5; SE = 4.6) as compared with either clinically localized prostate cancer ($P < 0.001$) or benign prostate tissue ($P < 0.0005$). There was no significant difference between the mean JAGGED1 staining between hormone naïve (score = 126.2; SE = 3.8) and hormone refractory metastatic prostate cancer (score = 129.1; SE = 9.7; $P = 0.79$). The findings were unchanged when minimum or maximum staining intensities were evaluated; data for minimum, mean and maximum intensity measurements are summarized with error bars with 95% confidence intervals in Figure 28. These data demonstrate an association between increased JAGGED1 expression and progression from localized to metastatic disease.

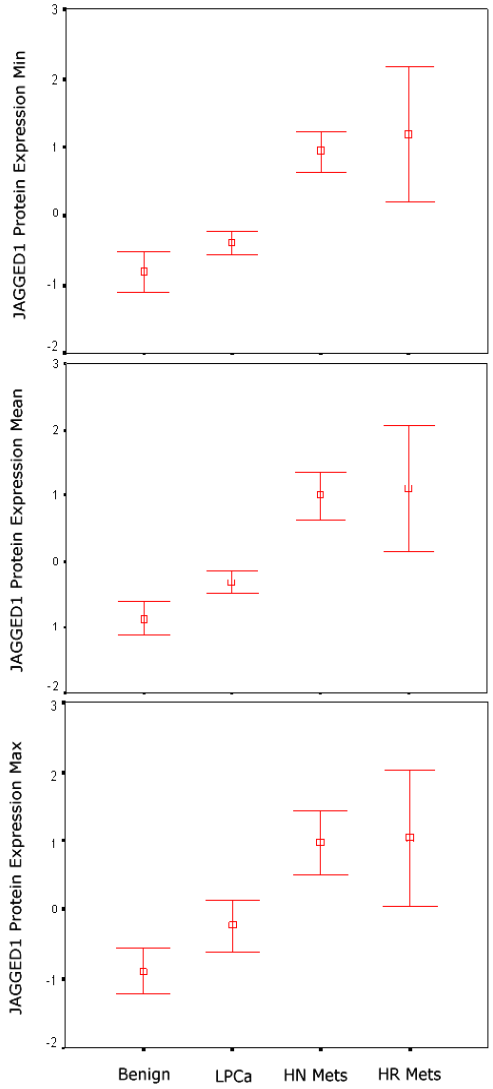


Figure 28 - JAGGED1 immunostaining in benign prostatic tissues, localized cancer, and metastatic hormone naïve and refractory can-

cer (summary). Confidence intervals (95%) show normalized minimum, mean and maximum protein intensity units of JAGGED1 as determine by quantitative evaluation of immunohistochemistry; bars, ±SE.

High JAGGED1 Expression Is Associated with Prostate Cancer Recurrence after Radical Prostatectomy for Clinically Localized Disease.

The increased expression of JAGGED1 in localized prostate cancer and the additional up-regulation in metastatic tumor suggested that JAGGED1 might be a useful tissue biomarker to facilitate differentiating indolent from more aggressive prostate cancer. To examine this possibility, an association was sought between JAGGED1 levels and prostate cancer recurrence, defined as an increase in prostate-specific antigen of 0.2 ng/ml after radical prostatectomy or the development of overt metastatic disease. JAGGED1 immunostaining was done on 3 previously validated tissue microarrays. Analysis was restricted to a subset of patients (n = 95) with clinically localized prostate cancer and 3 or more interpretable tissue cores, accordingly with [26]. Patient demographics, which are representative of the total tissue microarray cohort, are presented in Table 13. Among these 95 patients, 26 (27.4%) experienced prostate-specific antigen failure as of January 2004.

Table 13 - Clinical demographics of 95 men with localized prostate cancer (abbreviation: PSA, prostate-specific antigen).

		N %
Median Age (years)		59 (range 43 80)
Preoperative PSA (ng/ml)	<=4	17 (17.9%)
	and<10	56 (58.9%)
	>=10	22 (23.2%)
Digital Rectal Examination	negative	54 (56.8%)
	positive	41 (43.2%)
Gleason Score	<=6	35 (36.8%)
	7	55 (57.9%)
	>=8	5 (5.3%)
	3+4	10 (10.5%)
	4+3	45 (47.4%)

Surgical Margin Status	negative	64 (67.4%)
	positive	31 (32.6%)
Seminal Vesicle Invasion	negative	90 (94.7%)
	positive	5 (5.3%)
Extraprostatic Extension	negative	72 (75.8%)
	positive	23 (24.2%)
Median Tumor Dimension (cm)		1.4 (range 0.2-3.6)
PSA Failure (ng/ml)	no	69 (72.6%)
	yes	26 (27.4%)

The results of univariate analysis are shown in Table 14. Consistent with prior studies, Gleason score, preoperative prostate-specific antigen, extraprostatic extension, seminal vesicle invasion, positive surgical
margins, and digital rectal examination were all significantly associated with prostate-specific antigen failure at the univariate level. Interestingly, in univariate analysis, a strong association between prostate-specific antigen recurrence and high levels of JAGGED1 staining was found (relative risks of 2.55 with 1.55 intensity cutoff, and 3.05 with 1.8 intensity cutoff). It should be noted that significant *P*s were found for JAGGED1 only when maximum intensity values for each patient were used. The variation in staining across cores is consistent with the heterogeneous nature of prostate cancer and suggests that broad tissue sampling may be needed to maximize the predictive power of JAGGED1 staining. A Kaplan-Meier analysis depicting the association between JAGGED1 (cutoff of 1.8) and the probability of prostate-specific antigen free survival is shown in Figure 29.

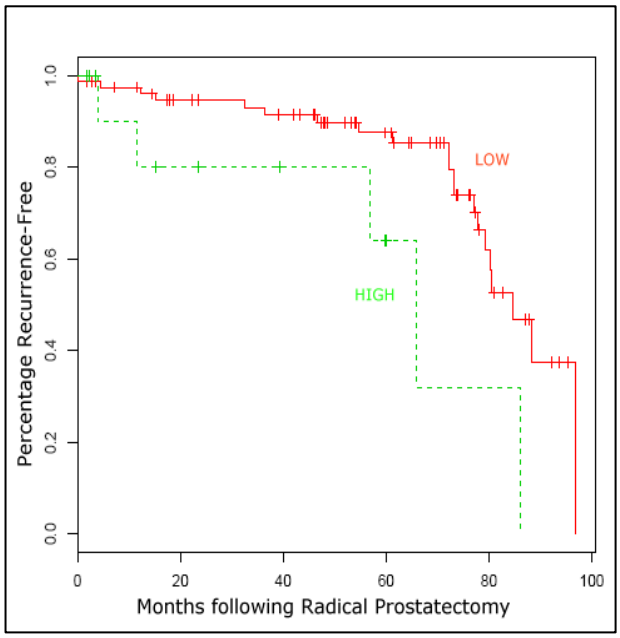


Figure 29 - Kaplan Meier analysis of individuals with clinically localized prostate cancer expressing high (cutoff of 1.8) or low levels of JAGGED1 (log rank test, p=0.02).

The best multivariate model predictive of prostate cancer specific recurrence after radical prostatectomy for clinically localized prostate cancer included extraprostatic extension (P = 0.0005; relative risk = 3.94), preoperative prostate-specific antigen (P = 0.042; relative risk = 1.97), and JAGGED1 (maximum intensity values were used; P = 0.016; relative risk = 3.51). Hence, high JAGGED1 protein level is a strong independent predictor of prostate cancer recurrence.

Table 14 - Parameters associated with prostate specific antigen recurrence following radical prostatectomy for clinically localized prostate cancer.

Univariate Analysis:	p-Value	Relative Risk	95% CI Lower Bound	Upper Bound
Gleason Score (2 cat, <7)	0.014	4.06	1.21	13.6
Preoperative PSA (3 cat)	0.025	2.43	1.31	4.51
Tumor Dimension	0.082	2.08	0.89	4.83
Extraprostatic Extension	0.0002	4.04	1.84	8.88
Seminal Vesicle Invasion	0.012	3.38	1.24	9.17
Positive Surgical Margins	0.0003	4.21	1.81	9.77
Digital Rectal Examination	0.038	2.29	1.02	5.15
JAGGED1 (cutoff=1.55) (Max)	0.041	2.55	1	6.48
JAGGED1 (cutoff=1.8) (Max)	0.021	3.05	1.12	8.31
Multivariable Analysis:				
Extraprostatic Extension	0.0005	3.94	1.84	8.40
Preoperative PSA	0.042	1.97	1.03	3.79
JAGGED1 (cutoff=1.8) (Max)	.016	3.51	1.26	9.76

7.3. Prostate Cancer Progression Profile

This study was conducted in collaboration with the Department of Pathology, Brigham and Women's Hospital, Boston, MA, Harvard Medical School, Boston, MA, the Children's Hospital Informatics Program, Children's Hospital, Boston MA, the Dana-Farber Cancer Institute, Boston, MA, the Department of Pathology and Urology, University of Michigan, Ann Arbor, MI, University Hospital of Ulm, Ulm, Germany.

The critical clinical question in prostate cancer research is to develop means of distinguishing aggressive from indolent prostate cancer. Expression array technology has lead to the development of discrete molecular signatures at transcript level [95], but the development of a robust signature to characterize aggressive prostate cancer has yet to be achieved.
Starting from a panel of genes accordingly dysregulated at transcript and protein levels [96], we studied a model to predict prostate progression. We performed *in situ* analyses of the correspondent proteins using a prostrate cancer progression TMA.

Interestingly this subset was able to distinguish men with clinically localized prostate cancer that were at highest risk to develop PSA-failure following surgery.
This study also demonstrates that cross platform models can lead to predictive models.

7.3.1 Material and method

Case Selection

We used a prostate cancer progression tissue microarray, described in [97,22]. It is composed of benign prostate tissue, localized prostate cancer, hormone naïve, and hormone refractory metastatic prostate cancer. These cases came from well-fixed radical prostatectomy, lymph node, and metastatic prostate cancer specimens from the University of Michigan (Ann Arbor, Michigan), the University Hospital Ulm (Ulm, Germany), and the rapid autopsy program (University of Michigan Specialized Program of Research Excellence (S.P.O.R.E.) in prostate cancer

Biomarkers for Immunohistochemisty

The majority of the biomarkers for this study were derived from a recent proteomics study. Refinement of this list of proteins included coordinate over or under expression by cDNA expression array analyses [98][99][100][101]. Antibodies against 41 proteins were optimized for *in situ* tissue evaluation by immunohistochemistry on archival formalin-fixed, paraffin-embedded material. The 41 biomarkers are presented in Table 15**.**

Semiautomated Quantitative Biomarkers analysis

Protein expression was evaluated by immunohistochemistry using an automated quantitative image analysis system, ACIS II (Chromavision Medical Systems, Inc, San Juan Capistrano, CA, USA). The ACIS II consists of a microscope with a computer controlled mechanical stage. Proprietary software is used to detect the brown stain intensity of the chromogen used for the immunohistochemical analysis and to compare this value to the intensity of the blue counter stain used as background. Intensity levels are

recorded as Intensity Units ranging from 0-255. Given the heterogeneity of the prostate tissue samples, the study pathologists used a computer-based selection tool to highlight areas within each 0.6 mm core for analysis. To account for this heterogeneity, we evaluated four tissue cores for each case.
In cases where less than three core where available, we substituted the data with the median value of the biomarker for this/that histologic subtype. The missing values can arise both from corrupted core sections (i.e., technically inadequate) and from a change of diagnosis. Missing values were present in the dataset 98 times in benign (13.3%), 130 times in localized prostate cancer (17.6%) and 6 times in the metastatic prostate cancer samples (0.8%). The change of histologic diagnosis was not a rare event and therefore supports the need to review all TMA cores (see chapter 5). As a pooling strategy we adapted the mean TMA core value for each patient.
The diagnosis of the selected area was recorded in the database as either benign, localized prostate cancer or metastatic prostate cancer. Cores with only stroma or non-diagnostic areas were excluded from further analysis. The hematoxylin and eosin stained images from this tissue array are available for review at a supplemental web site [102].

Statistical Analysis Clustering

Hierarchical agglomerative clustering both on samples and genes separately was carried out using Pearson correlation as similarity measure and average linkage method [103]. Clustering was performed using dChip software [104].

Linear Discriminant Analysis (LDA)
Fisher LDA was applied on the dataset of 41 genes to select genes [105,106] which discriminate among the diagnostics groups. Discriminant analysis is a model-free supervised approach (no assumption on data distribution are required), which uses both multivariate analysis of variance and discriminant procedure to

identify a linear combination of predictor variables that best characterizes the differences among the groups. LDA computes the so-called canonical variables (or canonical discriminant functions). The first canonical variable is the linear combination of the variables that maximizes the differences between the means of the groups (one dimension). The second canonical variable represents the maximum dispersion of the means in a direction that is orthogonal to the first canonical variable. The other canonical variables are generated in a similar manner. The number of canonical variables is given by the number of groups minus 1. Therefore in the current study with five tissue classes (i.e., benign, localized prostate cancer, hormone naïve metastatic prostate cancer, hormone refractory metastatic prostate cancer, and metastatic small cell prostate cancer), we obtained 4 canonical variables. When the two first canonical variables account for a large proportion of the variability in the dataset, a good graphical representation of the group differences can be obtained plotting the data along the first and the second canonical variables.

By applying a stepwise approach (adding and removing variables on variance evaluation), the most powerful subset of predicting variables can be defined. Stepwise selection begins by identifying the variable for which the means are most different and continues by adding the next best variable stepwise. Wilks'lambda method was used to control the entry or removal of predictor variables from the discriminant function. In discriminant analysis, prior probabilities were computed from group sizes. To measure the degree of success of the classification accuracy was evaluated. LDA was performed using R [107] and SPSS (SPSS Inc., Chicago, Il.).
We also applied separately feature selection algorithms to find most predictive marker set (diagnoses as class label); attribute selection involves searching through all possible combinations of attributes in the data to find which subset works best for class prediction. We used 'Best First', which iteratively adds attributes with the highest correlation with the class as long as there is not

already an attribute in the subset that has a higher correlation with the attribute in question; and ReliefFAttributeEval, which evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class. It can operate on both discrete and continuous class data. No significant differences were found.

Table 15 - Data description: Mean values and Confidence Intervals for the 41 markers and the 5 groups.

	BENIGN		LPCa		META		WAP		SM_CL	
	Mean	CI (95%)	Mean	CI (95%)	Mean	CI (95%)	Mean	CI (95%)	Mean	CI (95%)
ABP280	0.69	1.15	0.42	1.02	-1.61	0.71	-0.63	1.46	-0.57	0.02
AMACR-p504	-0.87	0.21	0.76	1.05	-0.20	2.05	0.64	2.82	-0.36	0.32
AR	0.19	1.92	0.37	1.57	-0.40	1.97	-0.44	2.14	-1.73	1.09
BM28	-0.61	0.55	0.11	1.41	0.07	2.98	0.54	1.90	2.27	1.64
BUB3	-0.06	1.20	0.42	1.10	-0.13	2.83	-0.82	3.35	0.27	1.73
CAMKK	-0.19	1.89	0.18	0.86	0.18	2.99	-0.05	3.08	-0.55	0.95
CASPASE3	1.02	0.85	-0.30	1.15	-0.17	1.94	-1.35	0.84	-1.03	0.68
CDK7	0.09	1.20	0.33	1.47	-0.58	2.39	0.19	3.00	-1.84	1.30
DYNAMIN	-0.07	1.74	0.73	1.05	-0.74	1.80	-0.83	2.25	0.25	2.74
E2F1	-0.06	1.78	0.07	1.22	0.46	2.70	-0.56	2.02	-0.12	5.72
E-CADHERIN	0.04	1.34	0.31	1.48	-0.66	2.00	0.64	2.13	-2.42	1.93
EXPORTIN	-0.02	1.27	0.24	1.44	-0.18	2.41	-0.21	3.96	-0.40	0.95
EZH2	-0.57	1.03	0.08	1.31	0.00	2.24	0.45	1.76	2.77	3.89
FAS	-0.31	1.94	0.08	1.58	0.22	2.54	0.47	2.40	-0.49	1.15
GAS7	-0.04	2.06	-0.09	1.49	0.05	2.27	0.24	2.83	0.12	3.13
GS28	0.00	1.46	0.28	0.91	0.34	0.53	-1.00	4.36	-0.53	2.19
ICBP90	-0.38	0.96	0.58	1.86	-0.47	2.30	-0.46	1.23	1.95	2.58
INTEGRIN	0.42	2.11	0.18	1.25	-1.34	1.31	0.31	1.42	-0.40	0.43
JAGGED1	-0.86	0.89	-0.30	0.47	1.11	0.98	1.23	2.20	1.15	0.74
JAM1	0.10	1.48	0.06	0.85	-1.24	2.43	0.75	1.37	1.56	3.48
KANADAPTIN	-0.58	1.29	0.23	1.12	-0.24	1.69	1.66	1.34	-1.54	1.25
KLF6	-0.07	1.40	-0.05	1.24	-0.62	2.66	1.11	2.60	-0.02	2.50
KRIP1	-0.45	1.59	0.33	1.58	0.02	1.66	-0.15	1.90	1.57	6.30
LAP2	-0.11	1.18	-0.02	1.14	-0.71	2.75	0.92	2.39	1.18	4.18

MCAM	-0.31	1.22	0.05	1.50	0.93	2.72	-0.16	2.38	-1.32	1.42
MIB1	-0.46	0.27	-0.17	0.52	-0.36	0.44	1.21	3.31	3.05	0.79
MTA1	-0.29	1.04	0.67	0.93	-0.99	2.33	0.38	2.91	-0.32	2.51
MUC1	-0.33	0.25	-0.30	1.28	-0.24	0.87	1.16	3.21	2.75	0.71
MYOSIN-VI	-0.56	1.11	0.24	1.73	0.58	2.10	0.49	2.66	-1.40	0.68
P27	0.43	1.73	0.20	1.78	-0.26	2.05	-1.30	1.06	0.04	0.11
P63-34beta-E12	1.29	0.46	-0.76	0.26	-0.77	0.24	-0.27	1.84	-0.34	0.48
PAXILLIN	0.00	1.45	0.32	2.30	-0.92	0.95	0.31	2.50	0.18	1.08
PLCLN	-0.54	0.83	0.63	1.21	0.05	3.16	-0.29	2.27	0.04	3.39
PSA	0.64	0.19	0.36	0.54	-0.27	1.84	-1.48	1.99	-2.57	0.07
RAB27	0.81	0.50	0.30	1.44	-0.95	1.78	-1.16	0.64	-1.62	0.27
RBBP	0.19	1.53	0.16	1.44	0.01	2.21	-0.75	3.52	-0.60	0.52
RIN1	0.79	1.32	0.35	1.36	-1.16	0.51	-1.13	0.46	-1.11	0.25
SAPKalpha	-0.24	1.63	0.05	0.99	-0.09	2.78	0.22	3.14	1.36	2.75
TPD52	-0.17	0.76	-0.04	1.35	1.22	2.42	-1.05	2.44	0.05	1.32
XIAP	-0.34	1.39	0.74	1.02	-0.90	1.58	0.58	2.35	-1.63	0.04
ZAG	0.44	1.11	0.34	2.17	-0.93	1.58	-0.42	1.76	-1.33	0.19

7.3.2 Results

Hierarchical Clustering Results

We performed high level analysis to check data quality present in the set of the 41 selected proteins. In particular, we investigated through hierarchical clustering if sufficient protein expression data could distinguish different states of prostate disease. Clustering was separately carried out on the samples and the 41 proteins. Highest levels of the sample tree (Figure 30) demonstrated good separation between aggressive prostate cancer states and clinically localized prostate cancer (LPCa). The clustering also reliably distinguished benign prostate tissue (BEN) from clinically localized prostate cancer (LPCa). Although the metastatic tumors clustered together, no clear subclusters were found for hormone naïve (META) and hormone refractory metastatic tumors (WAP), as demonstrated in Figure 30.Figure 30 - Protein Expression of 41 genes Selected for Differential Expression in Prostate Cancer

Progression. 1a. A “heat map” showing the relative protein expression of 41 genes selected as highly likely to demonstrate differential expression in prostate tissue samples along the spectrum of cancer progression. Protein expression was determined using antibodies directed against the various genes and measured by immunohistochemistry. The protein expression was determined using a semi-automated image analysis system (ASCIS II, Chromavision), which measures staining intensity along a continuous scale from 0 to 255. Low and high expression are depicted using light green and bright red, respectively. Hierarchical clustering of the samples demonstrates good but not perfect ability of this 41 gene panel to distinguish between the classes. 1b. Demonstrates that while some sub-types of metastatic prostate cancer (small cell cancer, upper figure, 200X original magnification) had discrete profiles, the clustering did not accurately distinguish between all of the hormone naïve and hormone refractory prostate tumor samples. Interestingly, a high-grade, clinically localized prostate cancer (Gleason pattern 4 prostate cancer, lower figure, 200X original magnification) was found to cluster more closely to the metastatic samples using this 41 gene profile. Two cases of metastatic small cell prostate cancer (SM_CL) clustered together (Figure 30b**,** top image). A sample of localized prostate cancer (LPCa_442-GL_7) was naturally grouped with the metastatic tumors. Although the overall Gleason score for this case was 7, the sample analyzed for this study depicted in Figure 30b demonstrates pure Gleason pattern 4 prostate cancer consistent with a high-grade tumor. When clustering genes based on samples, it is notably a group of seven genes over-expressed in benign tissues and under-expressed in aggressive cancer (p63, ZAG, ABP280, RAB27, RIN1, CASPASE3 and PSA). Extreme over and under expression of these proteins are present for aggressive cancer types (Figure 30, left side), supporting the hypothesis that the investigated set of markers might distinguish aggressive from indolent prostate cancer. The heatmap also suggests that some genes provide redundant or partially redundant information, as confirmed by descriptive statistics presented in Table 15.

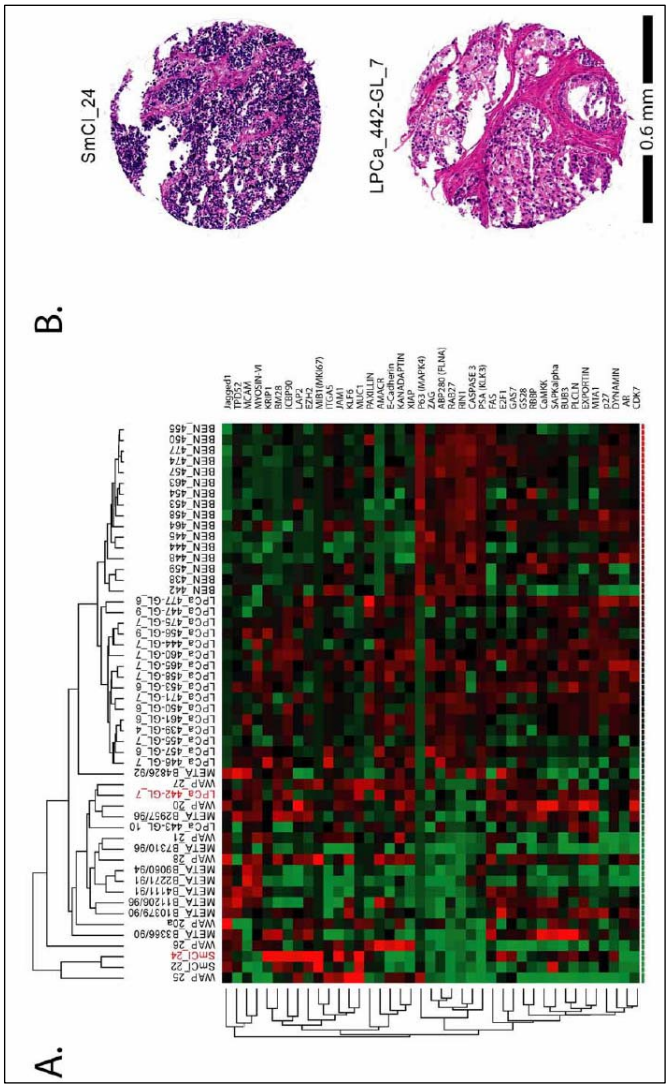


Figure 30 - Protein Expression of 41 genes Selected for Differential Expression in Prostate Cancer Progression. 1a. A "heat map" show-

ing the relative protein expression of 41 genes selected as highly likely to demonstrate differential expression in prostate tissue samples along the spectrum of cancer progression. Protein expression was determined using antibodies directed against the various genes and measured by immunohistochemistry. The protein expression was determined using a semi-automated image analysis system (ASCIS II, Chromavision), which measures staining intensity along a continuous scale from 0 to 255. Low and high expression are depicted using light green and bright red, respectively. Hierarchical clustering of the samples demonstrates good but not perfect ability of this 41 gene panel to distinguish between the classes. 1b. Demonstrates that while some sub-types of metastatic prostate cancer (small cell cancer, upper figure, 200X original magnification) had discrete profiles, the clustering did not accurately distinguish between all of the hormone naïve and hormone refractory prostate tumor samples. Interestingly, a high-grade, clinically localized prostate cancer (Gleason pattern 4 prostate cancer, lower figure, 200X original magnification) was found to cluster more closely to the metastatic samples using this 41 gene profile.

Linear Discriminant Analysis

In order to verify the discriminative power of the genes in terms of cancer progression and to identify gene profiles specific for localized prostate cancer and advanced prostate cancer, we developed a predictive model based on protein expression. Linear Discriminant Analysis (LDA) was applied to identify a linear combination of predictor variables that best characterizes the differences among the groups. Clear separation of the groups was found as depicted in Figure 31. The first and the second canonical variables cumulatively account for the 91.7% of the variance (68.1% and 23.6%, respectively). This result suggests that different groups (benign, localized cancer, hormone naïve metastases, hormone refractory metastases and small cells) are linearly separable in the gene space.

Stepwise linear discriminant analysis identified a set of 12 genes from the original set of 41 studied genes that best predicted tumor progression (Table 16) using 52 cases. Figure 31**b** represents the

cases along the first and second canonical components, which account for a cumulative variance of 87.9%.

The discriminative power of the 12 gene model was not decreased with respect to the 41 gene model, confirming also the redundancy of information provided by some genes; alternative subsets of genes from the 41 gene set could be selected. For example in the current study, after removing the 12 gene model from the original 41 genes, we can also identify good models with the remaining genes. The reason for associated gene expression patterns may be explained by the activation of similar molecular pathways or general processes such as proliferation or apoptosis.

Even though the model accuracy evaluated by cross validation (both using training and test sets 2/3 to 1/3 and leave-one-out) was very good, reliable performances should be assessed on a different larger dataset.

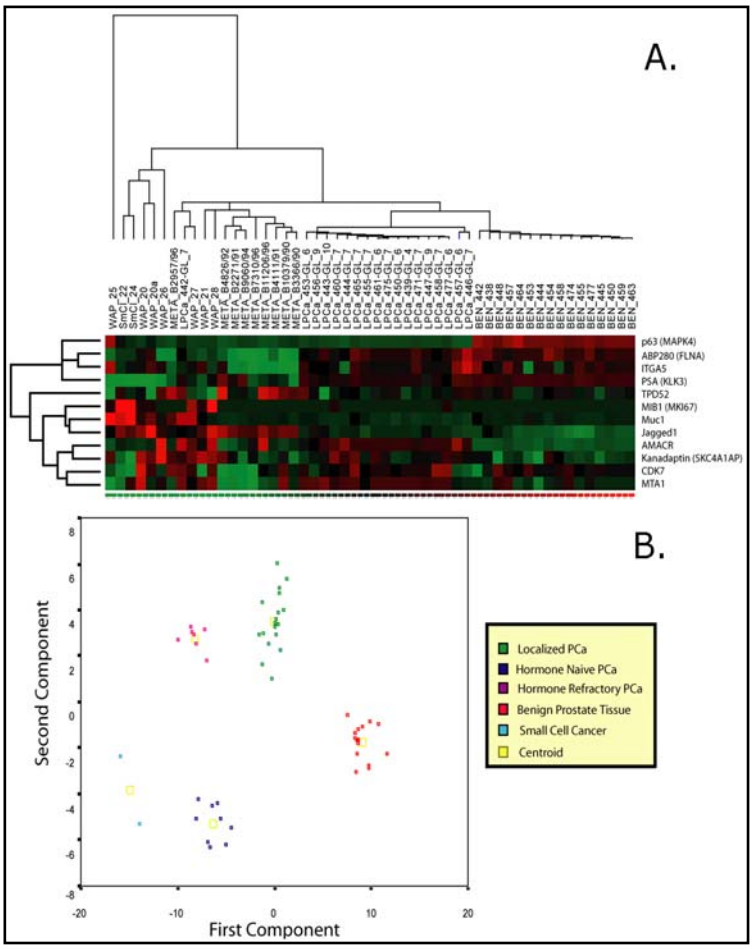


Figure 31 - Stepwise linear discriminant analysis (LDA) identified a group of 12 genes that best predict prostate cancer progression. (see Table 16). The heat map for these 12 genes is depicted in Figure 31a. Figure 31b represents the cases along the first and second canonical components of the LDA, which account for a cumulative variance of 87.9%. The discriminative power of the 12 gene model was not decreased with respect to the 41 gene model, confirming also the redundancy of information provided by some genes.

Table 16 - Stepwise LDA data: Fisher linear discriminant function coefficients (classification model).

	GROUP				
	BENIGN	LPCa	META	WAP	SM_CL
ABP280 (FLNA)	12,05	1,74	-16,81	-10,88	-8,60
AMACR	3,18	3,80	-10,10	-1,78	-14,03
CDK7	3,39	1,51	-2,94	-3,56	-19,28
ITGA5	-7,43	3,77	-2,73	9,44	0,27
JAGGED1	-10,04	-1,60	11,47	11,76	13,77
KANADAPTIN	-2,57	0,49	0,00	6,20	-3,41
MIB1 (MKI67)	-6,06	-4,08	5,75	8,76	32,84
MTA1	2,18	2,32	-8,73	0,02	-3,26
MUC1	-12,92	3,43	4,59	12,31	19,73
p63	41,51	-10,06	-24,94	-31,61	-49,72
PSA(KLK3)	11,82	2,26	-7,45	-14,78	-38,07
TPD52	2,07	-4,91	6,93	-3,88	12,23
(Constant)	-42,68	-9,22	-42,86	-47,72	-170,25

Expression Array Validation

In an approach described by Ramaswami et al. [8], the ability to distinguish between localized tumors likely to be aggressive from those that cured by surgery was successfully assesses at RNA expression level. Analysis was performed on a publicly available dataset of 80 localized prostate tumors [108] (data not provided).

7.4. Follow-up of the studies

The above described studies also served as starting points for further investigations, aimed to validate the results, to broader investigate the application of the analyzed proteins. They also encour-

aged the application of different experimental technique. Some details of the follow up of the studies are here reported.

M-CAM – This study has been submitted to Clinical Cancer Research. The next step will be the investigation of ß1 integrine expression. Experimental data [109] suggest that M-CAM could reduce ß1 integrine expression on cell membrane. A confirmation on our cohort would partially explain the poor prognosis associated with M-CAM expression.

Jagged1-This study was published in Cancer Research. The study has lead to the development of a funded pilot proposal by the Dana Farber Harvard Prostate Cancer Specialized Program of Research Excellence (SPORE). This work will try to validate the significance of Jagged1 as a prognostic indicator using a clinical dataset from Sweden with over 20 years clinical follow-up. Experimental studies will also explore the development of a urine Jagged1 diagnostic test for the presence of high-grade prostate cancer and the use of an inhibiter of Notch signaling to decrease prostate cancer cell growth in vitro.
S Santagata has been selected to receive the 2005 Stanley L. Robbins Memorial Research Fund Award for the project, "The role of JAGGED1 in prostate cancer progression".

Prostate Cancer Progression - The multigene model study will be presented as platform presentation at the USCAP (United States and Canadian Anatomic Pathology) meeting in San Antonio, Texas (2005) and at the Prostate Cancer SPORE meeting in Houston, Texas (2005). The Rubin laboratory is also in the process of developing a multiplex expression array test using a novel nanotechnology (Luminex) for the development of a tissue-based assay that could be applied on biopsy samples.

Chapter 8

8. Conclusions

Comprehensive analysis of protein functions in the context of their tissue environment will lead to the rapid identification, characterization, and development of novel diagnostic markers and therapeutic targets.

The investigation of molecular markers in surgical pathology has traditionally been accomplished by testing one marker at time on histological sections of tissues using immunohistochemistry or nucleic acid hybridization techniques. Each new marker had to be tested on multiple tissue samples. This was associated with the cost of each experiment, the variability between samples analyzed in the experimental run, and the added time associated with preparing and evaluating these samples.

The Tissue Microarray technology overcomes these limitations. Hundreds of individual tissue specimens can be arrayed on a single glass slide, which can be further probed and analyzed microscopically using the same standard laboratory methods (immunohistochemistry or *in situ* hybridization). These techniques allow simultaneous analysis of protein and RNA expression patterns in their tissue environment, as well as DNA profiling, making possible comparative studies.

High-throughput microscopic studies can be performed using less sample material, and a fraction of the reagents needed to visualize the target. Another important advantage is that TMAs produce uniform staining of multiple tissue sections on one slide, resulting in enhanced reproducibility. The conditions for each sample in a TMA are therefore uniform in contrast to the standard slide approach.

Immunohistochemistry is used to *in situ* determine the expression of proteins, which ultimately determine cell function. As immunohistochemistry is routinely used in pathology laboratories throughout the world, TMA studies have the potential to be more

translatable to a clinical application such as the development of diagnostic biomarkers or a potential to therapeutic target.
In TMA experiments, as with all high throughput techniques, high quality experimental data production is extremely important for the reliability of data analysis. This critical issue needs to be addressed on two levels, i. critical assessment of experimental design and organization and ii. reliability assessment of experimental data together with data pre-processing. In order to best deal with these two issues, a technological approach is advisable to properly manage data heterogeneity (biological, clinical and technical variables), data quantity and user diversity. This last aspect becomes even more crucial if data sharing occurs among and within research groups. Technological aspects may include automation to speed up data acquisition and evaluation that can be of great impact in overcoming TMA studies bottlenecks.

To efficiently and practically face the above mentioned problems a strong interdisciplinary effort must be taken. Lack of integration and uncontrolled data collection activities not only affect study results, but may also complicate the future ability to share data from the same patient or same protocols. This has in the past been the reason why experiments were often repeated unnecessarily on the sample patient cases with an associated financial and sample cost.

There are three critical issues related to the standard needs and requirements of TMA experiments: i. TMA studies may often involve considerable numbers of patients, heterogeneous data are involved, and different kind of users, often from different institutions, interact at different times; ii. TMA technology is prone to errors (association errors in designing the block array or evaluating a slide may easily occur) and iii. patients included in a TMA study (included in one block array) might be later on included in some other TMA studies or tissue based studies.

The technological solution we designed to address these requirements is a system that is i. patient centered and not experiment centered, ii. Work-flow oriented accounting for all phases of TMA experiments, iii. automated and iv. Web-based to allow for easy access and strong inter-institutional capabilities.
The latter characteristic is crucial when multiple institutions and/or department collaborate on specific studies. The patient centered solution gives great advantages in efficient data retrieval, both as TMA studies are usually based on several single experiments on the same cohort of patients and for new experiment design. In fact, this peculiarity enables the optimization of experimental designs by exploring tumor availability, easily constructing new defined experiments on sub-cohorts of patients by selecting cases and expression data across distinct TMA experiments, and further implementation to manage different experiment techniques directly sharing all the available data by simply adding new modules.
A key point that differentiates our approach from other solutions is the object recognition and ordering algorithm we integrated in the digital TMA acquisition system to automatically identify and associate each tissue sample with proper clinical information stored in the database. We favored the use of fully automated approach respect to semi-automated one to extremely speeding up the acquisition of single core section digital images by avoiding manual intervention. Regardless of the care and expertise used to create a TMA, some level of disorder with respect to the alignment of the tissue samples may arise, which forces the TMA technician to spend hours in appropriately aligning predefined grids, which has been used by most acquisition systems to date. Our system has the added advantage of being able to identify the tissue cores based on image recognition and knowledge of the TMA structure.
The TMABoost prototype system is in use since June 2003; up to now about 550 TMA experiments were supported by the system.

Most of the technological limits of our prototype system may be solved by industrial stage porting. Performances of the digital acquisition system can be bettered by interfacing different hardware. For example, scanner devices, which have better speed performances and are easy to use. An assessment of the performances of the automatic routines integrated in our system must be carried out using TMA slides prepared in other laboratories.

Up until now, data retrieval for analysis was always performed 'off-line'. Future work will focus on implementing on line data retrieval, to allow users to retrieve datasets through structured parametric queries.

Another aspect that must be implemented in the next future is the possibility to handle data ownership. This would include the possibility to share or separate data based on study and sample permissions. This becomes more critical as the database is used by multiple research groups at different institutions.

This thesis also addresses TMA data pre-processing issues.

Integrated frameworks for the collection and management of data are always of primary importance to ensure data quality. It applies in particular if different types of data and many data sources are involved at distinct time points. Technological approaches can dramatically reduce the lack of homogeneity and errors in raw data, by applying standardized consistency constraints and completeness controls. Nevertheless, they can only partially solve the problem and data preprocessing still remains a crucial phase of data analysis.

We identifying critical points and some solutions, tested on three TMA studies driven by different biological/clinical questions. We also evaluated applicability of automatic evaluations and addressed some of their limitations. Automation of TMA acquisition and evaluation process is highly desirable, but a fully automated approach is not currently possible and pathologist supervision is still required to obtain high quality data.

We also focused on a data analysis task. One particular problem to TMA studies is that there are usually multiple measures of the same protein for each patient, which are included in TMA dataset in order to account for tissue heterogeneity. We propose a novel classification model based on a Bayesian hierarchical approach, able to handle this kind of data uncertainty. This model was tested on two class problem simulated data, giving interesting results in terms of performances and output information. The next investigation step is validating the proposed model in a multi-feature classification framework and in other application domains.

This work highlighted many open issues in dealing with Tissue Microarray high throughput technology and proposed possible solutions. Some aspects concern general experimental requirements. In our experience, dealing with high throughput technology experiments imposes interdisciplinary effort. Drawbacks in the high throughput setting are mostly due to organizational issues, rather common in the bioinformatics field, and lack of integrated solutions to properly make sense of huge and diverse data. This is particularly true if multiple institutions collaborate on same studies, requiring standardized data sharing and standardized experimental protocols.
Another key point that we feel is becoming strategic, is that emerging studies will include different techniques in the same experimental process. This requires integrated, comprehensive and very flexible solutions, based on the management of all tissue samples for general experimental process. This approach enables to share previously studied cohorts or to include previously detected information and knowledge, taking advantage of well organized databases.

9. Acknowledgments

I would like to thank people who contributed to this work and supported me during this three years period. Among these a special acknowledgment is dedicated to Antonella Graiff, Andrea Sboner, Rossana Dell'Anna, Federica Ciocchetta and Paolo Traverso (ITC-irst, Trento I), Mattia Barbareschi and Paolo Dalla Palma (Pathology Department, Santa Chiara Hospital, Trento, I), Mario Stefanelli and the people of the Laboratory of Medical Informatics (University of Pavia, Pavia, I), Luigia Carlucci Aiello (Università di La Sapienza, Roma, I), Gianfranco Gensini (Università degli Studi di Firenze, Firenze, I), Dolores Di Vizio (Dana-Farber Cancer Institute, Boston, MA), Alberto Riva (CHIP, Children's Hospital, Boston, MA) and the people of the Rubin Laboratory (Brigham and Women's Hospital, Boston, MA).

10. Appendix A

To assess if and how initial calibration settings of the automatic system affect protein expression evaluations, we asked five pathologists to independently calibrate the system (inter-variability test). We used antibodies against MIB1 and ECAD. One pathologist set the calibration setting twice on one slide (intra-variability test).

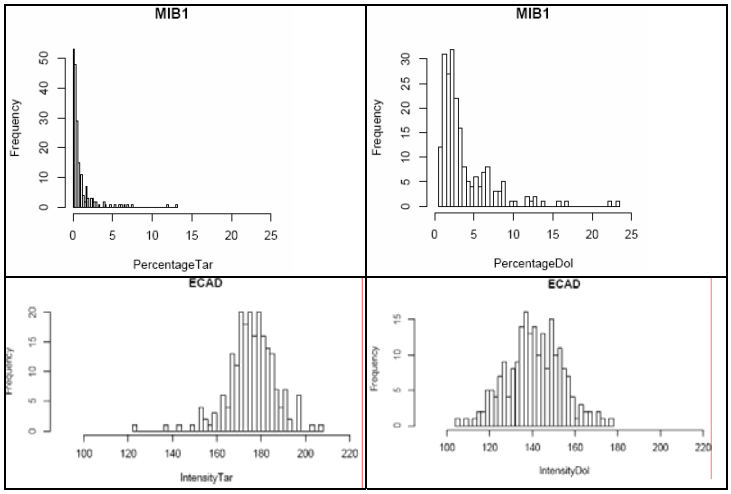


Figure 32 - Distribution of automatic evaluations for MIB1 percentages (first row) and ECAD intensities (second row). The two columns show data obtained with two different settings (two pathologists).

Output data distributions are shown in Figure 32. Variance tests (F test) among couples of evaluation data were performed. Data obtained for ECAD are shown in
Table 17.

Table 17 - F-Test results for ECAD. P values are reported where variance tests were statistically significant.

ECAD - Intensity

	Path1	Path 2	Path 3	Path 4	Path 5	Factory
Path 1		-	-	-	-	>0.05
Path 2			-	>0.05	-	-
Path 3				-	-	>0.05
Path 4					>0.05	-
Path 5						-
Factory						

We then looked for within group discriminative powers of each of the markers. We used mean pooled values among replicates. Wilcoxon test was applied. Results are listed in Table 18.

Population: Benign=18; Stroma=5; Adenocarcinoma=18; Metastatic PCA=9; WAP = 7; SmallCell=2.

Table 18 – Wilkoxon test results. P-values are reported where variance tests were statistically significant.

	BEN vs ADENO	ADENO vs META	ADENO vs WAP	ADENO vs (META + WAP)	META vs WAP	SM_CE vs WAP	STROMA vs BEN
	MIB1						
Path 1	<0.001	<0.05	<0.01	-	<0.05	-	<0.001
Path 2	<0.01	-	<0.01	-	<0.01	-	<0.01
Path 3	<0.001	<0.05	<0.01	-	<0.01	-	<0.05
Path 4	<0.001	<0.01	<0.05	-	<0.01	-	-
Path 5	<0.001	<0.05	<0.01	-	<0.01	-	-
Path 5b	<0.001	<0.01	-	-	<0.05	-	0.052
Factory	-	<0.01	-	-	<0.001	-	-
	ECAD						
Path 1	-	-	-	-	-	-	-
Path.2	-	-	-	-	-	-	-
Path 3	-	-	-	-	-	-	-
Path 4	-	<0.05	-	-	<0.05	-	-
Path 5	-	<0.05	-	-	-	-	-
Factory		-	-	-	-	-	-

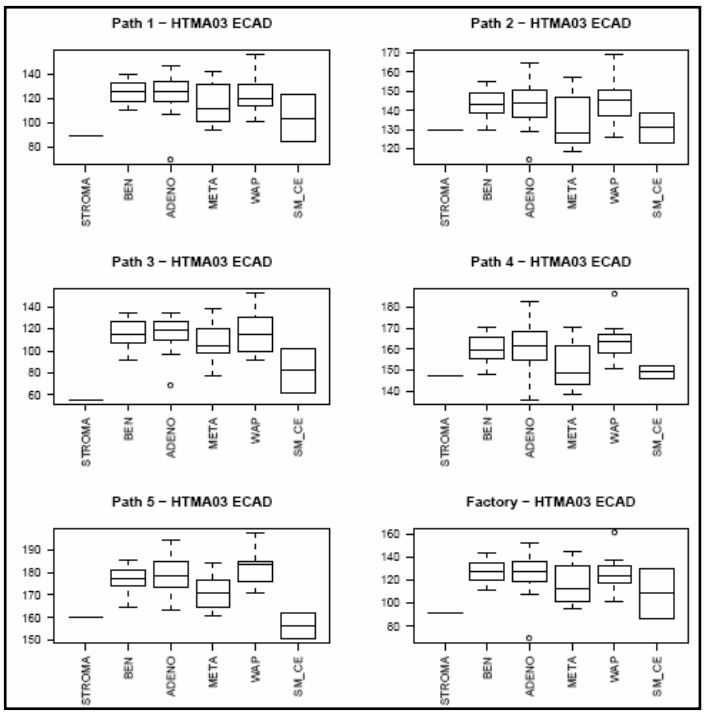


Figure 33 - ECAD intensities obtained with each setting set plotted against histotype groups. Cases used (mean value of core section replicates).

Based on the experiments we did, we observed that the distribution variance is not conserved when different settings are used. Output values are differently 'spread'.
The significance of mean differences between histotype groups is not always conserved. Intra-variability is also of some concern (see Path5 and Path5b on MIB1).
Calibration settings may affect experiment results.

Bibliography

[1] R. Dell'Anna, F. Demichelis, A. Graiff, A. Sboner. Cronache di laboratorio / Bioinformatica: Al servizio della medicina. Sapere, Edizioni Dedalo, 70, 5, 2004.
[2] F.Ciocchetta, R.Dell'Anna, F. Demichelis, A. Graiff, Andrea Sboner, and Linda Brodo. *Bioinformatics. A Review*, A.Graiff and C. Priami Eds. Franco Angeli - Fondazione Smith Kline, Milano, in press.
[3] N.M. Luscombe, D. Greenbaum, M.Gerstein. What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med 40(4):346-58. Review, 2001.
[4] C.M. Perou, T. Sorlie, M.B. Eisen, et al. Molecular portraits of human breast tumours. Nature, 17,406(6797):747-52, 2000.
[5] J. Lapointe, C. Li, J.P. Higgins, M. van de Rijn, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci USA, 20,101(3):811-6, 2004.
[6] T.R.Golub, D.K.Slonim, P.Tamayo, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286,531-537, 2000.
[7] L. Bullinger, K. Dohner, E. Bair, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med, 15,350(16):1617-28, 2004.
[8] S. Ramaswamy, K.N. Ross, E.S. Lander, T.R. Golub. A molecular signature of metastasis in primary solid tumors. Nat Genet, 33(1):49-54, 2003.
[9] D.R. Rhodes, M.G. Sanda, A.P. Otte, A.M. Chinnaiyan, M.A. Rubin. Multiplex biomarker approach for determining risk of prostate-specific antigen-defined recurrence of prostate cancer. J Natl Cancer Inst, 7,95(9):661-8, 2003.
[10] J.B. Welsh, L.M. Sapinoso, A.I. Su, et al. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Research, 15,61(16):5974-8, 2001.

[11] M. Schena, D. Shalon, R. Heller, A. Chai, P.O. Brown, R.W. Davis. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA; 93(20):10614-9, 1996.
[12] J. Kononen, L. Bubendorf, A. Kallioniemi, et al. Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nature Medicine, 4(7):844-847, 1998.
[13] L. Bubendorf, M. Kolmer, J. Kononen, et al. Hormone therapy failure in human prostate cancer: analysis by complementary DNA and tissue microarrays. J Natl Cancer Inst. 20,91(20):1758-64, 1999.
[14] Moch H, Schraml P, Bubendorf L, et al. High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. Am J Pathol. 54(4):981-6, 1999.
[15] T.J. Browne, M.S. Hirsch, G. Brodsky, W.R. Welch, M.F. Loda, M.A. Rubin. Prospective evaluation of AMACR (P504S) and basal cell markers in the assessment of routine prostate needle biopsy specimens. Hum Pathol, 35(12):1462-8, 2004.
[16] D. Zhang, M. Salto-Tellez, E. Do, T.C. Putti, E.S. Koay. Evaluation of HER-2/neu oncogene status in breast tumors on tissue microarrays. Hum Pathol, 34(4):362-8, 2003.
[17] P.J. Park, L. Tian, I.S. Kohane. Linking Expression Data with Patient Survival Times Using Partial Least Squares. Bioinformatics, 18:S120-7, 2002.
[18] E. Bair, R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol, 2(4):E108. Epub 2004.
[19] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, edited U. M Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/MIT Press, pp. 1-34, 1996.
[20] https://bioinfo.itc.it/TMA

[21] R. Dell'Anna, F. Demichelis, A. Sboner, M. Barbareschi. How to drive robotic microscope on TMA slides to measure biomarker expression: an ordering algorithm. CNIO Meeting - Tissue Microarray - Madrid, Sept 2003.
[22] M. A. Rubin, M. P. Zerkowski, R. L. Camp, R. Kuefer, M. D. Hofer, A. M. Chinnaiyan, D. L. Rimm. Quantitative determination of expression of the prostate cancer protein alpha-methylacyl-CoA racemase using automated quantitative analysis (AQUA): a novel paradigm for automated and continuous biomarker measurements. Am J Pathol, vol.164(3):831-40, 2004.
[23] J. J. Berman, M. E. Edgerton, and B. A. Friedman. The Tissue Microarray data exchange specification: A community based, open source tool for sharing tissue microarray data. BMC Med Inf and Decision Making, 3:5-13, 2003.
[24] R.L. Camp, L.A. Charette, D.L. Rimm. Validation of tissue microarray technology in breast carcinoma. Lab Invest, 80(12):1943-9, 2000.
[25] J. Torhorst, C. Bucher, J. Kononen.et al. Tissue Microarrays for Rapid Linking of Molecular Changes to Clinical Endpoints. American Journal of Pathology, 159:2249-2256, 2001.
[26] M.A. Rubin, R. Dunn, M Strawderman, K.J. Pienta. Tissue microarray sampling strategy for prostate cancer biomarker analysis. Am J Surg Pathol 26(3):312–319, 2002.
[27] R. Simon, G. Sauter. Tissue microarray (TMA) applications: implications for molecular medicine. Expert Rev Mol Med, 21:1-12, 2003.
[28] G.S. Bova, G. Parmigiani, J.I. Epstein, T. Wheeler, N.R. Mucci, M.A. Rubin. Web-based tissue microarray image data analysis: initial validation testing through prostate cancer Gleason grading. Hum Pathol, 32(4):417-27, 2001.
[29] L. Bubendorf, A. Nocito, H. Moch, G. Sauter. Tissue microarray (TMA) technology: miniaturized pathology archives for high-throughput in situ studies. J Pathol, 195(1):72-9, 2001.

[30] M.A. Rubin, S. Varambally, R. Beroukhim, et al. Overexpression, amplification, and androgen regulation of TPD52 in prostate cancer. Cancer Research, 64(11):3814-22, 2004.
[31] O. Ramuz, R. Bouabdallah, E. Devilard, N. Borie, A. Groulet-Martinec, V.J. Bardou, P. Brousset, F. Bertucci, F. Birg, D. Birnbaum, L. Xerri. Identification of TCL1A as an immunohistochemical marker of adverse outcome in diffuse large B-cell lymphomas. Int J Oncol, 26(1):151-7, 2005.
[32] C. Schmidt, M. Paraschar, W. Chen, J.F. Foran. Engeneering a peer-to-peer collaboratory for tissue microarray research. IEEE Proceedings of CLADE, 2004.
[33] A. M. De Marzo and H. Fedor. The Principles, Uses and Construction of Tissue Microarrays in Pathology Research. Gene Arrays and Tissue Arrays for Pathologists. At 92st USCAP Meeting, Washington D.C., 2003.
[34] C. L. Liu, W. Prapong, Y. Natkunam, A. Alizadeh, K. Montgomery, C. B. Gilks, M. van de Rijn, “Software Tools for High-Throughput Analysis and Archiving of Immunohistochemistry Staining data Obtained with Tissue Microarrays”, Am J Pathol, vol.161(5):1557-65, 2002.
[35] R. Shaknovich, A. Celestine, L. Yang, G. Cattoretti. Novel relational database for tissue microarray analysis. Arch Pathol, 127(4):492-4, 2003.
[36] S. Manley, N.R.Mucci, A.M. De Marzo, M.A. Rubin. Relational database structure to manage high-density tissue microarray data and images for pathology studies focusing on clinical outcome: the prostate specialized program of research excellence model. Am J Pathol, 62159(3):837-43, 2001.
[37] W. Chen, M. Reiss, D. J. Foran . A Prototype for Unsupervised Analysis of Tissue Microarrays for Cancer Research and Diagnostics. IEEE Trans. Information Technology in Biomedicine, 8(2):89-96, 2004.
[38] A.M. De Marzo, J.D. Morgan, B. Razzaque, C. White, J. Zimmerman, C.J. Bennett, H.L. Fedor, D. Faith. Tma-j: a set of

open source software tools to manage a multiple-organ, scalable, secure, multiple-user tissue microarray database. http://www.pathology.pitt.edu/apiii02/Sci-DeMarzo.htm, 2002.
[39] A. Hoos, C. Cordon-Cardo: Tissue microarray profiling of cancer specimens and cell lines: opportunities and limitations. Lab Invest, 81:1331-1338, 2001.
[40] X. Liu, V. Minin, Y.Huang, D.B. Seligson, and S. Horvath. Statistical methods for analysing tissue microarray data. J Biopharm Stat, 14(3):671-85, 2004.
[41] C. Bhattacharyya, L.R. Grate, M.I. Jordan, L. El Ghaoui, I.S. Mian. Robust sparse hyperplane classifiers: application to uncertain molecular profiling data. In press: *Journal of Computational Biology*, 2004.
[42] W.L. Buntine. Operations for learning with graphical models. J of Artificial Intelligence Research, 2:159-225, 1994.
[43] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin. Bayesian Data Analysis. Chapman & Hall/CRC, 2nd ed., 2004.
[44] P Sebastiani, E Gussoni, IS Kohane, MF Ramoni. Statistical Challenges in Functional Genomics. Statist. Sci. 18(1):33-70, 2003.
[45] R. Simon, M.Mirlacher, G. Sauter Tissue Microarray. Biotechniques, 36(1):325-35, 2004.
[46] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, edited U. M Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/MIT Press, pp. 1-34, 1996.
[47] N. Lavrac. Selected techniques for data mining in medicine Artificial Intelligence in Medicine, 16:3-23, 1999.
[48] K.J. Cios and G.W. Moore. Uniqueness of medical data mining. Artificial Intelligence in Medicine, 26(1,2), 2002.
[49] H. Battifora. The multitumor (sausage) tissue block: novel method for immunohistochemical antibody testing. Lab Invest, 55(2):244-8, 1986.

[50] O.P. Kallioniemi, U. Wagner, J. Kononen, G. Sauter. Tissue microarray technology for high-throughput molecular profiling of cancer. Hum Mol Genet, 10(7):657-62, 2001.
[51] D.L. Rimm, R.L. Camp, L.A. Charette, D. A. Olsen, E. Provost. Amplification of tissue by construction of tissue microarrays. Exp Mol Pathol, 70(3):255-64, 2001.
[52] C. Bucher, J. Torhorst, J. Kononen, P. Haas, J. Askaa, S.E. Godtfredsen, et al. Automated, high-throughput tissue microarray analysis for assessing the significance of HER-2 involvement in breast cancer [Abstract 2388]. Presented at the 36th Annual Meeting of the American Society of Clinical Oncology (ASCO) New Orleans, LA, May 20–23, 2000.
[53] C. Bucher, J. Torhorst, L. Bubendorf, P. Schraml, J. Kononen, H. Moch, et al. Tissue microarrays ("tissue chips") for high-throughput cancer genetics: linking molecular changes to clinical endpoints. Am J Hum Genet, 65(Suppl):A10, 1999.
[54] N.R. Mucci, G. Akdas, S. Manely, M.A. Rubin. Neuroendocrine expression in metastatic prostate cancer: evaluation of high throughput tissue microarrays to detect heterogeneous protein expression. Hum Pathol, 31(4):406-14, 2000. Erratum in: Hum Pathol 31(6):778, 2000.
[55] R. Simon, G. Sauter. Tissue microarray (TMA) applications: implications for molecular medicine. Expert Rev Mol Med, 21:1-12, 2003.
[56] A. Warford, W. Howat, J. McCafferty. Expression profiling by high-throughput immunohistochemistry. J Immunol Methods, 290(1-2):81-92, 2004.
[57] R.L. Camp, G. G. Chung, D. L. Rimm. Automated subcellular localization and quantification of protein expression in tissue microarray. Nature Med, 8(11):1323-1327, 2002.
[58] M.S. Fejzo, D.J. Slamon. Frozen tumor tissue microarray technology for analysis of tumor RNA, DNA, and proteins. Am J Pathol, 159(5):1645-50, 2001.

[59] I. Jacobson, G. Booch, J. Rumbaugh. The Unified Software Development Process. Addison-Wesley Pub Co; 1st edition (February 4, 1999).
[60] F Demichelis, A Sboner, R Dell'Anna, J Santi, M Barbareschi, A Graiff. A web-based system for the management of Tissue Microarray: fostering biomedical research. Proceedings of Medinfo, San Francisco, CA, (CD):1571, 2004.
[61] M. Barbareschi, F. Demichelis, S. Forti, P. Dalla Palma. Digital Pathology: Science Fiction?. Int J Surg Pathol, 8(4):261-263, 2000.
[62] F. Demichelis, M. Barbareschi, P. Dalla Palma, S. Forti. The virtual case: a new method to completely digitize cytological and histological slides. Virchows Arch, 441(2):159-64, 2002.
[63] SQL Server Magazine, SQL Server Magazine: The XML files,http://msdn.microsoft.com/library/default.asp?url=/library /en-us/dnsqlmag2k/html/TheXMLFiles.asp, 2000.
[64] E.F. Codd Derivability, redundancy, and consistency of relations stored in large data banks. Research Report RJ599, IBM, August 1969.
[65] E.F. Codd. The Relational Model for Database Management Version 2. Addison-Wesley, 1991.
[66] A. Lucenti et al. Web-based automated tissue microarray system analysis of Her2/neu expression and prognostic significance in a monoinstitutional series of 436 cases using Herceptest®. American Society of Clinical Oncology, Memorial Convention Center, New Orleans, Louisiana, June 5 - 8, 2004.
[67] A. Sboner, M. Barbareschi, R. Dell'Anna, and F. Demichelis. Large scale TMA experiments: automation and data management. BITS - Bioinformatics Italian Society Meeting, Padova, March 2004.
[68] J. Quackenbush. Computational analysis of microarray data. Nature Rev Genet, 2:418–427, 2001.

[69] S. Dudoit, J. Fridlyand and T.P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Amer. Statist. Assoc, 97:77–87, 2002.
[70] D.R. Cox. Regression models and life-tables. J of Royal Stat Soc. B, 34:187-220, 1972.
[71] T.G. Clark, M.J. Bradburn, S.B. Love and D.G. Altman. Survival Analysis Part IV: Further concepts and methods in survival analysis. British Journal of Cancer, 89, 781-786, 2003.
[72] M.W. Kattan. Judging new markers by their ability to improve predictive accuracy. Edotprial. J Natl Cancer Inst, 7;95(9):634-5, 2003.
[73] F.E. Harrel, K.L. Lee, R.M. Califf, D.B. Pryor, and R.A. Rosati. Regression modelling strategies for improved prognostic prediction. Statistics in Medicine, 3(2),143:152, 1984.
[74] N.J.D Nagelkerke. A note on a general de¯nition of the coe±cient of determination. Biometrika, 78, 691:692, 1991.
[75] P.J. Park, Gene Expression Data and Survival Analysis, in Shoemaker, JS and Lin, SM, Editors, Methods of Microarray Data Analysis IV, Springer, New York, 2004.
[76] H. Li and J. Gui. Partial Cox regression analysis for high-dimensional microarray gene expression data. Bioinformatics, 20 Suppl. 1:i208–i215, 2004.
[77] Chuaqui RF, Bonner RF, Best CJ et al. Post-analysis follow-up and validation of microarray experiments. Nat Genet, 2 Suppl:509-14, 2002.
[78] A.K. Jain, M.N. Murty, P.J. Flynn. Data Clustering: A Review. ACM Computing Surveys, Vol. 31, No. 3, 1999.
[79] K.A. Fleming. Evidence-based pathology. J Pathol, 179(2):127-8, 1996.
[80] D. Kahneman, P. Slovic, A. Tversky. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, 1982.
[81] S. Paik, J. Bryant, E. Tan-Chiu, et al. Real-world performance of HER2 testing--National Surgical Adjuvant Breast and Bowel Project experience. J Natl Cancer Inst, 94(11):852-4, 2002.

[82] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and Regression Trees, Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, CA, 1984.
[83] J. Friedman and N. Fisher. Bump hunting in high dimensional data. Statistical Computing, 9:123-143, 1999.
[84] P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifiers. In AAAI-92, 1992.
[85] T.M. Mitchell. Machine Learning, McGraw-Hill, New York, 1997.
[86] I.M. Shih, The role of CD146 (Mel-CAM) in biology and pathology. J Pathol, 189(1):4-11. 1999.
[87] C. Doglioni, C. Chiarelli, E. Macri, A.P. Dei Tos, E. Meggiolaro, P. Dalla Palma, M. Barbareschi. Cyclin D3 expression in normal, reactive and neoplastic tissues. J Pathol, 185(2):159-66, 1998.
[88] M. Barbareschi, L. Pecciarini, M.G. Cangi, et al. p63, a p53 homologue, is a selective nuclear marker of myoepithelial cells of the human breast. Am J Surg Pathol, 25(8):1054-60, 2001.
[89] N. Bardin, F. Anfosso, J.M. Masse, et al. Identification of CD146 as a component of the endothelial junction involved in the control of cell-cell cohesion. Blood, 15,98(13):3677-84, 2001.
[90] M. Barbareschi, O. Caffo, S. Veronese, et al. Bcl-2 and p53 expression in node-negative breast carcinoma: a study with long-term follow-up. Hum Pathol, 27(11):1149-55, 1996.
[91] D.B. Martin, D.R. Gifford, M.E. Wright, et al. Quantitative proteomic analysis of proteins released by neoplastic prostate epithelium. Cancer Research, 64:347–55, 2004.
[92] S. Santagata, F. Demichelis, A. Riva, et al. Jagged1 Expression Is Associated With Prostate Cancer Metastasis And Recurrence. Cancer Research, 64(19):6854-7, 2004.
[93] S.M. Dhanasekaran, T.R. Barrette, D. Ghosh, et al. Delineation of prognostic biomarkers in prostate cancer. Nature, 412:822–6, 2001.

[94] E.E. Perrone, C. Theoharis, N.R. Mucci, et al. Tissue microarray assessment of prostate cancer tumor proliferation in African-American and white men. J Natl Cancer Inst, 92:937–9, 2000.
[95] S.M. Dhanasekaran, T.R. Barrette, D. Ghosh, et al. Delineation of prognostic biomarkers in prostate cancer. Nature, 23; 412(6849):822-6, 2001.
[96] S. Varambally, J. Yu, B. Laxman, et al. Integrative Proteomic and Genomic Analysis of Prostate Cancer Progression, submitted.
[97] M.A. Rubin, M. Putzi, N. Mucci, D.C. Smith, K. Wojno, S. Korenchuk, et al. Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. Clin Cancer Res, 6(3):1038-45, 2000.
[98] D.R. Rhodes, T.R. Barrette, M.A. Rubin, D. Ghosh, A.M. Chinnaiyan. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res, 62(15):4427-33, 2002.
[99] D.R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, et al. ONCOMINE: A Cancer Microarray Database and Integrated Data-Mining Platform. Neoplasia, 6(1):1-6, 2004.
[100] D.R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci USA, 101(25):9309-14, 2004.
[101] B.S. Stein, S. Vangore, R.O. Petersen. Immunoperoxidase localization of prostatic antigens. Comparison of primary and metastatic sites. Urology, 24(2):146-52, 1984.
[102] http://rubinlab.tch.harvard.edu.
[103] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A, 95(25):14863-8, 1998.

[104] C. Li, W.H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A, 98(1):31-6, 2001.
[105] R.A. Johnson, D.W. Wichern. Applied multivariate statistical analysis. 5th ed. Upper Saddle River, N.J.: Prentice Hall; 2002.
[106] H. Jiang, Y. Deng, H.S. Chen, L. Tao, Q. Sha, J. Chen, et al. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics, 5(1):81, 2004.
[107] R. Ihaka, R. Gentleman. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics, 5(3):299-314, 1996.
[108] G.V. Glinsky, A.B. Glinskii, A.J. Stephenson, R.M. Hoffman, W.L. Gerald. Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest, 113(6):913-23, 2004.
[109] S. Alais, N. Allioli, C. Pujades, J.L. Duband, O. Vainio, B.A. Imhof, D. Dunon. HEMCAM/CD146 downregulates cell surface expression of beta1 integrins. J Cell Sci, 114(Pt 10):1847-59, 2001.