COMPUTATIONAL BIOMEDICINE: A MULTIDISCIPLINARY CROSSROADS

Yves Moreau
Kathleen Marchal
Janick Mathys

January 2002

Katholieke Universiteit Leuven
Department of Electrical Engineering
ESAT-SCD (SISTA)

Kasteelpark Arenberg 10
B-3001 Heverlee, Belgium

Email: {Yves.Moreau,Kathleen.Marchal,Janick.Mathys}@esat.kuleuven.ac.be

With the collaboration of

Prof. Bart De Moor
Dr. Mik Staes
Stein Aerts
Peter Antal
Bert Coessens
Tijl De Bie
Frank De Smet
Kristof Engelen
Geert Fannes
Patrick Glenisson
Qizheng Sheng
Gert Thijs

Ten geleide…

'Waarom gebeurt er onderzoek rond *bio-informatica* in een departement *elektrotechniek* ?' Sinds een drietal jaren, krijg ik deze vraag regelmatig voor de voeten geworpen.

Een eerste laconieke reactie zou kunnen zijn: 'Waarom ook niet?' De veeleer historische naam 'Elektrotechniek' dekt immers niet langer de lading. Bekijk maar eens de onderzoektopics op de website van ons departement (www.esat.kuleuven.ac.be)! Maar bij een tweede, meer doordachte reactie, leg je dan toch maar uit dat bio-informatica een *inter- en multi-disciplinair* gebeuren is, en dat het er uiteindelijk niet toe doet in welk departement precies je zo'n onderzoeksgroep opzet. Je voegt er aan toe dat de klassieke, 'verticale', universitaire departementen historisch gegroeid zijn rond domeinen die lang geleden duidelijk waren afgebakend, zoals biologie, scheikunde, boerderijbouw, werktuigkunde of pakweg sinologie; maar dat tegelijkertijd diezelfde universitaire structuren minder geschikt zijn voor de uitbouw van 'horizontale', multidisciplinaire onderzoeksplatformen. Laat staan om deze te financieren. Hoort bio-informatica thuis in een onderzoekscommissie biologie, genetica, oncologie, statistiek of informatietechnologie? Het lijkt soms wel op een theologische discussie uit de Middeleeuwen!

Duidelijk is echter dat wetenschap en technologie in toenemende mate een belangrijke impact hebben op onze kennismaatschappij. Duidelijk ook is dat steeds meer maatschappelijk themata zich 'horizontaal' ontwikkelen, met hun vele dimensies in wetenschappelijk onderzoek, technologie, ethische, juridische en socio-democratische problematieken. Voorbeelden zijn legio: milieu en duurzame ontwikkeling, de interconnectiviteit en interactiviteit van de informatiemaatschappij, verkeer en mobiliteit, en alles wat met gezondheidszorg heeft te maken.

Specifiek voor bio-informatica is het veelbetekenend dat we bij vele collegae uit andere departementen, bij de bio-ingenieurs, de biologen, de statistici, de collegae uit de geneeskunde en farmacologie, maar ook bij de juristen, enorme weerklank krijgen op ons initiatief, en dat we er in toenemende mate in slagen om ook gezamenlijke onderzoeksprojecten te formuleren.

Een onderzoeksgroep rond bio-informatica dus, in de onderzoeksafdeling ESAT-SCD, waar we sinds jaar en dag onderzoek verrichten rond numerieke lineaire algebra, statistiek, systeemtheorie, algoritmen voor geavanceerde regeltechniek, optimalisatie-algoritmen, neurale netwerken, kortweg alles wat je zou kunnen catalogeren onder de noemer van 'mathematical engineering' (zie www.esat.kuleuven.ac.be/sista-cosic-docarch). Deze inzichten en technieken worden toegepast in tal van toepassingsdomeinen, zoals telecommunicatie, biomedische signaalverwerking, modelgebaseerde industriële procescontrole, en met de exponentiële groei van numerieke databanken van vandaag de dag, ook in 'numerieke data-mining'. Deze inzichten werden in het recente verleden gevaloriseerd in verschillende spin-off bedrijven (www.ipcos.be, www.data4s.com, www.tmleuven.be).

Bio-informatica als onderzoeksdiscipline is geheid op wetenschappelijke en technologische doorbraken op drie niveaus:

- Belangrijke 'milestones' in de moderne moleculaire biotechnologie ('de dubbele helix', PCR, de ontrafeling van steeds meer genomen, waaronder dat van de mens,…);
- Grote vooruitgang in computertechnologie, zowel op het gebied van hardware ('De wet van Moore') als op het gebied van software (databanken, interconnectiviteit, search engines, enz…);
- Grote vooruitgang in sensormethodologieën en data-acquisitie om biologische en genetische processen experimenteel te monitoren ('DNA-chips', microarrays,…);

Ons onderzoeksteam bio-i is dan ook multidisciplinair (zie www.esat.kuleuven.ac.be/~dna/BioI/index.html): 1 postdoc burgerlijk ingenieur, 3 postdocs bio-

ingenieur, en een tiental doctoraatstudenten, waaronder 2 burgerlijk ingenieurs, 3 bio-ingenieurs, 2 fysici, 2 masters of AI en zowaar een burgerlijk ingenieur die ook dokter in de geneeskunde is. Dit is een uniek team, dat bovendien regelmatig geruggesteund wordt door verschillende andere onderzoekers van ESAT-SCD en elders. Bijzonder fier ook zijn we op onze prille wetenschappelijke resultaten (www.esat.kuleuven.ac.be/~dna/BioI/Publications.html), met reeds op korte tijd toppublicaties in wereldtijdschriften, waaronder Nature Genetics (Brazma et al., 2001). En voor wat betreft het onderwijs, staan we mee aan de wieg van een 'Masters in bioinformatics', die sinds dit academiejaar georganiseerd wordt (www.esat.kuleuven.ac.be/sista/GGS/).

Als de tweede helft van de 20ste eeuw het ontstaan heeft gezien van de informatietechnologie en de moleculaire biologie, dan zal de eerste helft van de 21ste eeuw misschien wel herinnerd worden als het tijdperk van de integratie van bio- en informatietechnologie. Het voorliggend werkstuk wordt ingediend door drie van onze post-docs (Kathleen Marchal, Janick Mathys en Yves Moreau), met de medewerking van vele doctoraat- en eindwerkstudenten. Het reflecteert hun recente bijdragen in dit prille onderzoeksdomein. We hopen dan ook van harte dat het inspirerend mag werken op toekomstige generaties van onderzoekers in de b(l)oeiende nieuwe wereld van de bio-informatica.

Bart De Moor
ESAT-SCD, K.U.Leuven

31 januari 2002

Table of contents

Ten geleide… 2

Table of contents 4

I. Introduction 7

I.1. The road to computational biomedicine 7

I.2. Overview of this work 8

I.3. Computational techniques and software tools 10

I.4. Economic and social context 11

I.5. Acknowledgments 13

II. Machine learning in medical decision support 14

II.1. Ovarian and endometrial cancer 14

II.1.1. Endometrial cancer 15

II.1.2. Ovarian cancer 15

II.2. Techniques for prediction of tumor malignancy 15

II.2.1. Logistic regression 15

II.2.2. Artificial neural networks 15

II.2.3. Bayesian networks 17

II.3. Design of the data 18

II.3.1. Endometrial cancer data 18

II.3.2. Ovarian cancer data 19

II.3.2.1. The IOTA protocol 20

II.3.2.2. The IOTA database 20

II.3.2.3. The IOTA web site 20

II.3.2.4. A general case reporting tool 21

II.4. Development of the models 22

II.4.1. Logistic regression for the prediction the depth of invasion in endometrial carcinomas 22

II.4.2. ANN models for the discrimination of malignant and benign adnexal masses 22

II.4.3. Bayesian networks for ovarian tumor diagnosis 22

II.5. Evaluation of the models 25

II.5.1. Evaluation of the ANN model 25

II.5.2. Evaluation of the logistic regression model 27

II.5.3. Evaluation of Bayesian networks 28

II.6. Discussion 29

II.7. Annotated Bayesian networks 29

II.7.1. Semantic network representation of Annotated Bayesian Networks 30

II.7.2. Annotated Bayesian Networks in information management 31

II.7.3. Annotated Bayesian Networks in decision support 32

II.7.4. Information retrieval using Annotated Bayesian Networks 33

II.8. Conclusion 34

III. Microarrays in oncology: from disease management to fundamental research 35

III.1. Microarrays 36

III.2. Necessity for a laboratory information management system 38

III.3. Preprocessing of simple experiment design (black & white) 40
III.3.1. Sources of noise 40
III.3.2. Mathematical transformation of the raw data: need for a log transformation 42
III.3.3. Filtering data 43
III.3.4. Ratio approach 43
III.3.5. Analysis of variance 44

III.4. Microarrays and disease management 45
III.4.1. Molecular fingerprinting of malignancies 46
III.4.2. Development of a data-mining framework 46
III.4.2.1. Feature selection 47
III.4.2.2. Class prediction 48
III.4.2.3. Cluster analysis of patients or microarray experiments 50

III.5. Conclusion 51

IV. From expression to regulation 53

IV.1. Clustering gene expression profiles 53
IV.1.1. Adaptive quality-based clustering 54
IV.1.2. Clustering gene expression in the mitotic cell cycle of yeast 57

IV.2. Using functional annotations of genes: Gene Ontology 59

IV.3. Whole-genome scan for genes regulated by a known transcription factor 63
IV.3.1. Introduction 63
IV.3.1.1. Zinc finger proteins 63
IV.3.1.2. Transcription factors 64
IV.3.1.3. Structure of genes 65
IV.3.1.4. Transcription 65
IV.3.1.5. Role of transcription factors in cancer 65
IV.3.2. Methodology 66
IV.3.2.1. Study of the occurrences of PLAG1 binding motifs in the promoter region of a known target 66
IV.3.2.2. Search for the motif in the promoters of genes upregulated by PLAG1 66
IV.3.2.3. Search for multiple occurrences of the PLAG1-binding motif in known human promoters 67
IV.3.2.4. Full genome scan for multiple occurrences of the PLAG1-binding motif 67
IV.3.3. Results 68
IV.3.3.1. Occurrence of the motif in the promoter of genes upregulated by PLAG1 68
IV.3.3.2. Occurrence of the motif in known human promoters 69
IV.3.3.3. Full genome scan 69
IV.3.4. Future research 69

IV.4. Motif finding in sets of co-expressed genes 70

IV.5. First steps towards integration of tools 72

IV.6. Conclusion 74

V. Towards integrative genomics 75

V.1. Genetic network inference: a complex problem 75
V.1.1. High-level methodology 76

V.2. Experimental setup 77

V.3. Development of a genetic network inference method in the Bayesian context 77

V.4. Development of an ontology-based knowledge system of yeast and Salmonellae for acquisition of prior knowledge *78*
V.4.1. Benefits of the knowledge system 79
V.4.2. Ontologies 79
V.4.3. Benefits of ontologies 80
V.4.4. Methodology 81
V.5. From local to global repositories: MIAME *82*
V.6. Conclusion *83*

VI. General conclusion **84**

Reference List **86**

I. Introduction

This work is at the crossroads of medical informatics and bioinformatics. It addresses how computation, statistics, and information technology affect research at the interface between medicine and biology. It illustrates how these methodologies run through the different subdisciplines of medicine and biology and, even more importantly, how they link these disciplines. A wide range of techniques is presented: from standards for data storage, to machine-learning approaches for medical decision support and to data mining of gene activities. Further, the application of these techniques is demonstrated at the hand of real-life examples (many of them relating to oncology) – for example, the diagnosis of ovarian tumors or the detection of genes involved in tumor development.

I.1. The road to computational biomedicine

This work stands where two historical trends collide: the exponential increase in computing power and the exponential increase in biomolecular data. These trends are perhaps most vividly exemplified by the fact that molecular biology and chemistry are overtaking fluid dynamics, weather forecasting, and virtual nuclear testing as the most power-hungry computing applications. As an example, Celera of Rockville, MD deployed for the assembly of the human genome one of the top commercial supercomputers (a cluster of 800 processors with 70 terabytes of storage). Furthermore, it is now collaborating with Sandia National Labs (a U.S. government research laboratory previously specialized in nuclear weapons) for the development of a supercomputer of the next generation. This supercomputer will run at 100 teraflops (10^{14} floating-point operations per second), which will make it 8 times faster than today's fastest civilian supercomputer. To quote Paul Robinson, president of Sandia: "We in the nuclear weapons community felt for many years nothing could be more complex than nuclear physics, but I'm now convinced nothing beats the complexity of biological science." Also, traditional computer companies such as Silicon Graphics, Compaq and IBM are strongly moving towards applications in the life sciences. As an additional example, IBM is currently working on the development of BlueGene, a 1 petaflop (10^{15} flops) supercomputer dedicated to the study of protein folding. Yet, not only computing speed is essential, but also the handling of massive quantities of data. For example, genome sequence information is doubling every 18 months (which coincidentally parallels the evolution in computing speed predicted by Moore's law). Furthermore, the distributed character of biological information (several hundreds of databases of genomic information are maintained by experts all over the world) and the underlying biological complexity make it a bigger challenge to handle this type of information than, let us say, financial transactions or weather measurements. Finally, new experimental techniques (such as the microarray technology discussed later in this work) will shortly produce data at a rate higher than Moore's law thereby putting computing environment under increasing strain. In fact, experts often mention a future production of biological information of 100 gigabytes per day at leading research facilities. For these reasons, electrical engineering and computer science have an essential role in addressing the immense challenge that the life sciences offer for the 21st century.

In medicine, the increasing prevalence of computerized information (medical imaging, electronic patient records, automation of clinical studies) considerably enhances the further progress of medicine as a data-driven evidence-based science, alongside its empirical tradition. As a result, medicine is developing tighter and tighter links to engineering, computer science, and statistics. In biology, faced to the flood of data generated by high-throughput genomics (Human Genome Project, Arabidopsis Genome Initiative, microarrays, Single Nucleotide Polymorphism Initiative and so on), biologists have a pressing need for support, guidance, and collaboration for the analysis of their data. The importance of data management and analysis cannot be underestimated,

as it has become a main bottleneck in molecular biology (which itself is a driving force of the pharmaceutical and biotechnological sectors).

Information technology provides a practical platform for a better integration of the different biological and medical disciplines, both for practice and research. As a result, we witness the convergence of the many disciplines relating to the application of computation and information technology to biology and medicine, such as (together with some examples):

- Medical information systems (electronic patient records)
- Biostatistics (design and analysis of clinical studies and clinical trials)
- Medical decision-support systems (diagnosis assistance and critiquing)
- Biomedical image analysis (radiography, nuclear magnetic resonance)
- Biomedical signal processing (electroencephalography, electrocardiography, and also as an essential initial step for image analysis)
- Biomedical systems and control (intelligent prostheses, intelligent drug delivery devices)
- Statistical genetics and epidemiology (gene mapping, single nucleotide polymorphism analysis)
- Computational structural biology (prediction of protein structure from sequence)
- Biological databases and information technology (gene and protein databases)
- Bioinformatics and computational biology (statistical data analysis strategies for molecular biology and *in silico* biology)

We call Computational Biomedicine the general discipline resulting from this convergence. This evolution is a long-term trend that will continue over several decades. The work we present here shows how this convergence is actually happening by bringing together elements from medical information systems, biostatistics, medical decision support, biological information technology, and bioinformatics for a series of medical and biological applications.

I.2. Overview of this work

This manuscript integrates several topics from medical informatics and bioinformatics, most of them with applications in oncology. The two major topics are the development of methodologies for decision support in medical diagnosis and for the analysis of microarray data (measurements of gene activities) for medical and biological applications. Two themes also progress through the manuscript. The first one is the movement from empirical medicine to a medicine more deeply rooted in statistics (evidence-based medicine) and in biology and chemistry. The second one is a transition from applied medical cases (namely, the prediction of malignancy in tumors) to fundamental biological problems (such as the study of gene activity during the yeast cell cycle or the discovery of genomic motifs related to gene regulation). The more fundamental problems serve obviously as the foundation for future applied work.

In *Chapter 2*, we first introduce a medical task on which we have been working for several years in collaboration with the Department of Obstetrics and Gynecology of the University Hospitals of Leuven: the preoperative discrimination of ovarian tumors. The goal there is to predict whether an ovarian tumor is benign or malignant on the basis of patient information (e.g., age, number of pregnancies,…) and of ultrasonographic measurements (such as the size and shape of the tumor). We also introduce a second similar task: the prediction of the malignancy of endometrial carcinoma from ultrasonographic measurements. After introducing the ovarian tumor problem, we describe how such data has been collected and explain the web application we have developed for this purpose. This tool is currently used in an international study lead by the Department of Obstetrics and Gynecology of the University Hospitals of Leuven to collect about 1000 case reports per year, which makes it the largest database on this topic in the world. We then go on to describe several statistical models for the prediction of malignancy that we developed on an earlier database of such records. First, we introduce the basic logistic regression. Next, we present neural networks, which are more complex but also have a better performance. We describe two

different neural networks that we have designed for this task. We then discuss the performance of the different models.

Yet, even though black-box models (such as logistic regression and neural networks) make quite good classifiers, they lack interpretability and the possibility of incorporating expert knowledge in the decision-support system. For these reasons, we introduce the methodology of belief networks, which are probabilistic models of the data distribution. These models provide a principled way of handling uncertainty and of modeling the relationship between the different variables present in the task. We demonstrate the performance of belief networks as classifiers for the classification of ovarian tumors and also show how to build these models by first incorporating expert knowledge in the model and then refining the model with statistical data. We extend the framework of Bayesian networks with a new methodology, called Annotated Bayesian Networks, that allows the tracking of the large volumes of documentation necessary for building complex Bayesian networks. Beyond information management, we show that this methodology is also efficient in decision support and information retrieval.

In *Chapter 3*, we introduce microarray technology, which is one of the recent technologies that are having a major influence on research in molecular biology. Microarrays are miniaturized devices that measure the activities of thousands of genes in a sample in a single experiment. This technology is going to contribute to a better management of cancer by profiling the molecular response of individual malignancies. Maybe more importantly, because of their high-throughput nature, microarrays have become a method of choice for the study of gene function and regulation on a global genomic scale. They are invaluable for unraveling the networks of regulation that control the dynamic behavior of genes. This is an essential element in understanding the network of interaction between genes, which is the central goal of genomics.

First, we introduce the basics of microarray technology. With such a high-throughput technology, it becomes immediately clear that data storage is a significant challenge. We briefly describe how a Laboratory Information Management System keeps track of all the steps necessary for the deployment of microarray technology. Next, we discuss the preprocessing of the raw data from microarray experiments. These experiments have a high level of noise because of inherent technological limitations. It is therefore essential to clean up the data to obtain reliable measurements. To this end, we present two classes of techniques: normalization and analysis of variance (ANOVA).

Second, as an illustration of the principles of microarray data analysis and to make clear how microarray technology fits within the clinical themes from the previous chapter, we show how to use microarray data in building models for disease management in oncology. At the hand of microarray data from two types of leukemia, we illustrate the different tasks in this area: (1) the selection of the features most relevant to some clinical outcome, (2) the prediction of the clinical outcome using statistical models similar to those presented in the previous chapter, and (3) the discovery of classes of malignancies at the molecular level (which could be different from the current medical classification and could provide more insight into the behavior of the malignancy) by clustering algorithms.

In *Chapter 4*, we then switch to the analysis of microarray data for elucidating genomic processes. We present an integrated methodology that combines microarray and genome sequence data to discover which patterns control gene expression at the sequence level. The first part of the methodology is the identification of groups of genes that show similar behavior. The underlying assumption is that, among the genes showing the same behavior (we say that they are co-expressed), some may share the same control pattern at the sequence level (we say that they are then coregulated). The identification of relevant groups can be done by studying known groups of genes (we discuss how nomenclature efforts, such as the Gene Ontology, help to define such groups) or by clustering algorithms. We present a new clustering algorithm, called adaptive quality-based clustering, which overcomes some of the limitations of classical clustering

algorithms. We demonstrate the power of this method on microarray data from the yeast cell cycle. The second part of the methodology is the identification of which short motifs in the genomic sequence are likely to control gene expression. This identification can be done either by screening the groups of co-expressed genes for known motifs and detecting which motifs are common to many genes in the group or by trying to build the motif pattern from scratch using statistical methods. For scanning for a known motif, we discussed the identification of new targets of the PLAG1 transcription factor involved in benign tumors of the salivary gland. Further, we present the method of Gibbs sampling for motif finding that we have developed. We have implemented the different strategies just described in a web application called INCLUSive (Integrated Clustering and motif Sampling, http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html).

In *Chapter 5*, we unveil our ongoing research. We show how the Bayesian network methodology presented in Chapter 2 can be extended to the interpretation of microarray data. Bayesian networks give a structured framework to model the interaction between genes and handling the uncertainty inherent to microarray measurements; they therefore make the inference of genetic networks from microarray data possible. However, limitations in the measurements available make it essential to incorporate *a priori* knowledge. Again, Bayesian networks are the ideal platform for the integration of heterogeneous (and sometimes contradictory) sources of information. Furthermore, we describe a general framework that integrates all the methods we present and many more into a knowledge pipeline for molecular biology. There we discuss further how ontologies are central to large-scale knowledge management in genomics. For microarrays in particular, our group at ESAT-SCD has been involved in the development of microarray standards within the Microarray Gene Expression Database (MGED) consortium (http://www.mged.org), namely in the development of the Minimal Information About a Microarray Experiment (MIAME) standard, which we describe here.

Finally, we briefly present the conclusions of this work.

I.3. Computational techniques and software tools

A wide range of computational techniques is necessary to tackle the different problems addressed in this work. To give a bird's eye view on the computational side of our work, we list these techniques together with their corresponding application:

- Elementary statistical tests
  - Performance assessment of predictive models
  - Feature selection in microarray data
  - Detection of overrepresentation of known motifs in genomic sequences
- Logistic regression
  - Prediction of malignancy in ovarian tumors and endometrial carcinoma
- Neural networks
  - Prediction of malignancy in ovarian tumors
- Bayesian networks
  - Prediction of malignancy in ovarian tumors
  - Genetic network inference from microarray data
- Principal component analysis
  - Analysis of microarray data
- Clustering
  - Analysis of microarray data
    - K-means clustering
    - Adaptive quality-based clustering
- Gibbs sampling for missing data
  - Motif finding in genomic sequences

Further, many of the methods used in this work have been implemented as standalone or web-based tools:

- IOTA web application for collection and validation from patient case reports in ultrasonography
- Environment for probabilistic modeling with Bayesian networks and Annotated Bayesian Networks
- Laboratory information management system for microarray design
- MAGOSeq web application for exploration of microarray data using gene ontologies and for sequence exploration
- INCLUSive web application for clustering of microarray data and motif discovery in genomic sequences
- Application for genetic network inference with Bayesian networks

The INCLUSive application is publicly available (http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html) and the IOTA application is available to the 15 research groups participating in the IOTA study. Our laboratory information management system for microarrays served as prototype for the system currently in use at the Microarray Facility of the Flemish Institute for Biotechnology.

I.4. Economic and social context

At the onset of the 21st century – and notwithstanding their astonishing successes in the previous century and major new achievements lying within arms reach, both the healthcare and the pharmaceutical industry face an uncertain future. The victim of their own success, they are now threatened by cost containment measures by governments across the world and by rising quality demands (as witnessed by the steady pressure from malpractice lawsuits, the high vigilance of consumer groups, and the increasingly stringent requirements of drug approval agencies in the form of larger clinical trials). The growth of the healthcare market slows down because of the economic limits healthcare has reached. Total healthcare spending in the U.S. (both pharmaceutical and medical) shot from $250 billion to $700 billion from 1980 to 1990, but then increased more slowly to reach $1.300 billion in 2000. In percentage of the U.S. gross economic product (GDP), healthcare went from 9% to 12% of the GDP from 1980 to 1990, but settled at 13% of the GDP in 2000. Simply put, society spends as much as it can on healthcare and drugs – but not more.

Furthermore, the pharmaceutical industry faces an increase in drug development costs combined with a decrease in return per drug[1]. Between 1988 and 1998, the U.S. pharmaceutical industry has seen its R&D costs soar from $6.5 billion a year to $21 billion a year. The trend in the European industry is exactly the same and the situation in 2002 is unchanged. According to a recent study[2], the cost per drug, adjusted for inflation, has risen from $300 million to $800 million between 1987 and 2000. In 1998, R&D costs represented 17% of sales revenues while they represented only 10% in the 1980s. At the same time, the growth of the drug market is slowing: while the market was growing at an average of 11% a year in the 1980s, it was growing only at 6% a year in 1997, with no improvement in sight because of general cost containment measures. Increasing competition caused by lowering barriers to entry in drug development and generic drugs gnaws further away at the revenues. Because drug development is a high-risk activity, the pharmaceutical industry must return significant profits to its investors. Indeed, only one in ten drugs entering preclinical development will make it to the public. The development time is long: typically between 7 and 10 years. The cost is huge: on average $600-$800 million for each successfully developed drug. Furthermore, the return is uncertain: on the one hand, 90% of all

[1] PriceWaterhouseCoopers, *Pharma 2005: An Industrial Revolution in R&D*, 1998.
[2] Tufts Center for the Study of Drug Development, http://www.tufts.edu/med/csdd/Nov30CostStudyPressRelease.html.

drugs earn less than $200 million a year; on the other hand, Pfizer's cholesterol lowering drug Lipitor earned $6 billion in 2001. Between 1993 and 1998, the top 20 pharmaceutical companies delivered a return of 20% a year (capital growth and dividends). However, the current perspective is much more dire and, for the coming ten years, it will be a major challenge for most companies to deliver even half of this return. Worse, if current trends cannot be turned around, the pharmaceutical industry might become unattractive to investors and the whole industry could stall, bringing the development of new drugs to a crawl.

Following the landmark of the Human Genome Project, genomics and bioinformatics are revolutionizing the industry, promising fast and cost-effective development of new drugs. The fight against the impending menace will take place on many fronts: genomics, chemo- and bioinformatics, virtual testing, pharmacogenomics, and a tighter integration of the discovery, development, and trial phases. The completion of the human genome and the advent of the post-genomic era promise a flood of new drug targets to the pharmaceutical industry and a bonanza of biomarkers to the diagnostics industry. Current drugs use only about 500 different molecular targets while it is estimated that genomics and proteomics could eventually provide between 5.000 and 10.000 targets. The question is then moving from discovering targets to predicting which targets have the best potential. As mentioned before, the amount of data produced by new techniques from molecular biology and chemistry is exploding. Chemoinformatics and bioinformatics will be essential to mining these mountains of data. Data handling and analysis will cover the whole drug development processes, tackling questions such as which genes are involved in a pathology, which compounds are likely to show toxic effects, or which patients could present rare side effects. An especially exciting trend is the emerging combination of genomics and bioinformatics for the development of *in silico* models of cells, organs, or even patients. By building extensive mathematical models of biological processes on the basis of genomics measurements, it will become possible to prescreen targets and compounds *in silico*. This improves the quality of the candidates that enter the development phase, thereby significantly reducing development costs. Another trend is that of pharmacogenomics, which links drug response to the specific genetic profile of an individual. By identifying those individuals who present rare side effects as having specific genetic variations, it will be possible to rescue some drugs that fail late in the development process (and for which the investment has been maximal) by linking their use to a genetic screening of the patient. Similarly, drugs that fail because they are not active on a sufficiently large portion of the patients could be rescued in some cases (e.g., anti-cancer drugs). Finally, a tighter integration of the whole process (for example, by feeding back genomic patient information into the discovery process) will also increase the efficiency of the development process.

Clearly, for both the healthcare and the pharmaceutical industry, the only way out is the way forward, which means delivering better medical procedures and better drugs more efficiently and more safely, together with targeting problems for which there is a high social demand (chronic and degenerative diseases (such as AIDS, Alzheimer's disease, or arthritis), cardiovascular and metabolic diseases, or cancer). This goal implies an integrated view of the patient in the healthcare process and an intimate understanding of pathologies from the socioeconomic and psychological levels to the genetic and molecular levels. Our work contributes humbly to the technical side of this social endeavor. It addresses questions in oncology stretching from the clinic to the wet lab, such as collecting data from patients for clinical studies, predicting diagnosis from clinical variables, moving new methods from molecular biology towards clinical practice, and studying basic processes in biology as a foundation to medical research. Recurring themes in our work are the focus on a more personalized medicine and the development of computational models that achieve a better understanding of the biological processes at hand, in particular pathologies. The articulation of our different projects into the coherent framework of computational biomedicine participates to the development of the integrated and personalized medicine of the 21st century.

Finally, let us not forget that medicine is for people. To reach its full effect, technical work, like ours, must be embedded in the social, economic, legal, and psychological dimensions of our society. We must make better medicine available to the largest number. Finally, we must insist that medical care is much more than a technical act – empathy and communication are just as essential.

I.5. Acknowledgements

Yves Moreau, Kathleen Marchal, and Janick Mathys thank all the members of the ESAT-SCD (SISTA) Bioinformatics team for their essential contribution to this work: Prof. Bart De Moor, Stein Aerts, Peter Antal, Bert Coessens, Tijl De Bie, Frank De Smet, Kristof Engelen, Geert Fannes, Patrick Glenisson, Qizheng Sheng, Mik Staes, Gert Thijs. They thank Jos De Brabanter for much assistance with statistics and Prof. Joos Vandewalle and Prof. Sabine Van Huffel for their support. They also thank the many people with whom they have been extensively collaborating in the past few years: Prof. Dirk Timmerman, Prof. Ignace Vergote, Prof. Pierre Rouzé, Stéphane Rombouts, Magali Lescot, Prof. Yves Van de Peer, Dr. Paul Van Hummelen, Tom Bogaert, Prof. Wim Van de Ven, Dr. Marianne Voz, Karen Hensen, Prof. Bart De Strooper, Dr. Mike Davrowski, Hannelore Denys, Prof. Jos Vanderleyden, Sigrid De Keersmaecker, Pieter Monsieurs, Prof. Johan Thevelein, Prof. Joris Winderickx, Johnny Roosen, and Dr. Torik Ayoubi.

Yves Moreau
Kathleen Marchal
Janick Mathys

Leuven,
31 januari 2002

II. Machine learning in medical decision support

In this chapter, we look into several statistical and machine learning methods for medical decision support, specifically for tumor diagnosis in oncology. The preoperative discrimination between malignant and benign tumors is a crucial issue in gynecology. Since the beginning of 1997, our research group has been closely cooperating with Prof. Dirk Timmerman and Prof. Ignace Vergote of the Department of Obstetrics and Gynecology of the University Hospitals of Leuven. This collaboration has led to the setup of an international study with 15 centers from Europe and the U.S., the International Ovarian Tumor Analysis (IOTA) consortium. The goal of the study is to collect the largest database of ultrasonographic case reports from patients with ovarian tumors in the world (about 1000 cases per year) and to develop predictive models based on statistics and artificial intelligence for the preoperative assessment of such tumors. In a pilot project, based on the data of 300 patients, stepwise multivariate logistic regression analysis was performed to select relevant parameters from the preliminary data. In a next step, various logistic regression models were developed and tested. Logistic regression analysis, based on simple parameters (such as menopausal status, serum level of the tumor marker CA125, a score based on the intensity of blood supply in the tumor and the presence of papillary structures), could be used to preoperatively discriminate benign and malignant ovarian masses in a reliable way. Also artificial neural networks (ANNs), based on simple parameters (such as age of the patient, CA125 serum level, and some morphologic features), were trained to reliably predict malignancy. In a prospective study, the neural network performed significantly better than the widely used 'Risk of Malignancy Index'. A first statistical study of the difference in performance indicated that the developed ANN models were still very close to the logistic regression model, in other words that the possible model complexity of the ANNs was not fully exploited. This means that many hard-to-classify (i.e., close to the decision boundary) examples are needed to train ANNs and to significantly enhance their global performance. Next to the growing number of collected patient data, a large amount of medical background knowledge is also available. These two sources of information result in two different modeling strategies. From the medical background knowledge, leading medical experts can construct various discrimination models, which are then tuned and tested by qualitative observations (knowledge models). From clinical measurements, various statistical models can be developed (data-driven models), such as the logistic regression and artificial neural networks (ANNs).

In this chapter, we describe the use of both approaches for the discrimination between malignant and benign tumors in the adnexa (ovaries, fallopian tubes) and the endometrium. Simple statistical models, logistic regression models, and ANNs (which all predict the malignancy of a tumor based on collected observations) are described and compared. For the combination of prior knowledge and observations, we give an overview of the Bayesian network models that were developed in our research group. For all these techniques we summarize potential applications and report the performance of such models in ovarian and/or endometrial cancer diagnosis.

II.1. Ovarian and endometrial cancer

Endometrial and ovarian tumors, the latter belong to the group of the adnexal masses, are the result of uncontrolled tissue growth in the endometrium and the ovaries. Normally, cells divide only when additional cells are required for normal body function. However, at certain times the controls that regulate the division of the cell are lost. This results in unordered growth of more and more cells into a mass that is termed a tumor. It should be stressed that not all tumors are a cause of cancer. The tumors can be either benign or malignant. Malignant means that the tumor cells spread to other tissues (metastasis). Benign tumors are usually not life threatening, as they do not cause metastasis. Borderline ovarian tumors are a third class of ovarian tumors that have the

cytological features of malignancy but do not invade the ovarian stroma and have a very good prognosis.

II.1.1. Endometrial cancer

Cancer or carcinoma of the endometrium is the most common female pelvic malignancy. The endometrium is the inner lining of the uterus. Most of the tumors are confined to the uterus at diagnosis and can be cured. Nevertheless, endometrial carcinoma is still the seventh leading cause of death from cancer in women. The morphology of the tumor and the blood flow (in the uterine arteries and in the tumor itself) are visualized by transvaginal sonography (gray scale) and color Doppler imaging (CDI).

II.1.2. Ovarian cancer

Ovarian malignancies represent the greatest challenge among the gynecologic cancers. In the absence of a family history of ovarian cancer, women have a lifetime risk of ovarian cancer of about 1/70. Early detection is of primary importance, since currently more than two-thirds of the patients are diagnosed with advance disease.
The morphology of and blood flow in the tumor is determined on the basis of ultrasound images. In this way, observations are made about morphologic characteristics such as the locularity, papillation and about the vascularization of the mass (e.g., the resistance index). Additional diagnostic information is obtained by measuring the serum levels of tumor markers such as CA125.
The risk of malignancy index (RMI) was introduced as a combination of the several types of data (Tingulstad et al., 1996). The gynecologist determines scores for the menopausal status and for the morphology of the mass. The serum CA125 level is measured and all three values are multiplied. If the result (RMI) exceeds a fixed threshold, the tumor is predicted to be malignant. In most cases, however, the patient has to undergo surgery to obtain tissue samples for pathology.

II.2. Techniques for prediction of tumor malignancy

Several techniques were applied for the development of mathematical models to help the gynecologist make a correct assessment in patients with ovarian or endometrial tumors. The following techniques were used: logistic regression, ANNs, and Bayesian networks.

II.2.1. Logistic regression

One of the machine learning techniques used for predicting the malignancy of a tumor is logistic regression. Logistic regression is a variation of ordinary regression, used when the outcome is restricted to two values (malignant or benign). It produces a formula that predicts the probability of the outcome as a function of the independent variables. A special s-shaped curve is fitted on the data by taking the linear regression, which produces any y value between minus and plus infinity, and transforming it with the function

$$p = 1 / (1 + \exp(y)),$$

which produces values of p between 0 (benign) and 1 (malignant).

II.2.2. Artificial neural networks

ANNs are networks of interconnected processing elements (nodes), inspired by the connectivity of neurons in the brain (Haykin S., 1994). They can be used for identification of patterns in data, by exposing them to large amounts of similar data containing inputs with their corresponding

outputs. Metaphorically one could say that the network learns from experience just as we do. The structure of the ANNs that were used for prediction of the malignancy of ovarian tumors is outlined in Figure 2.1 and 2.2.

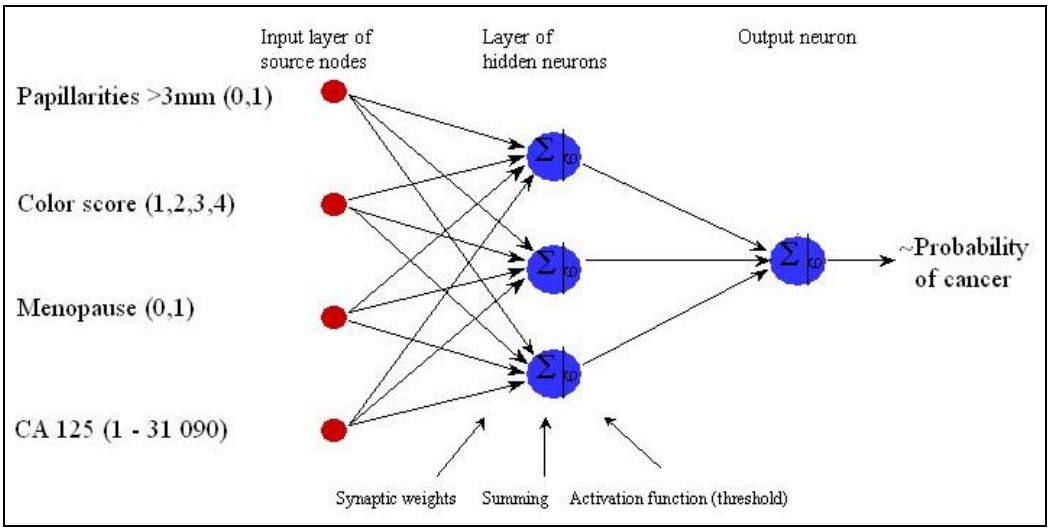


Figure 2.1. Structure, input variables, and output of the first ANN used for prediction of malignancy of adnexal masses. (Timmerman et al., 2000)

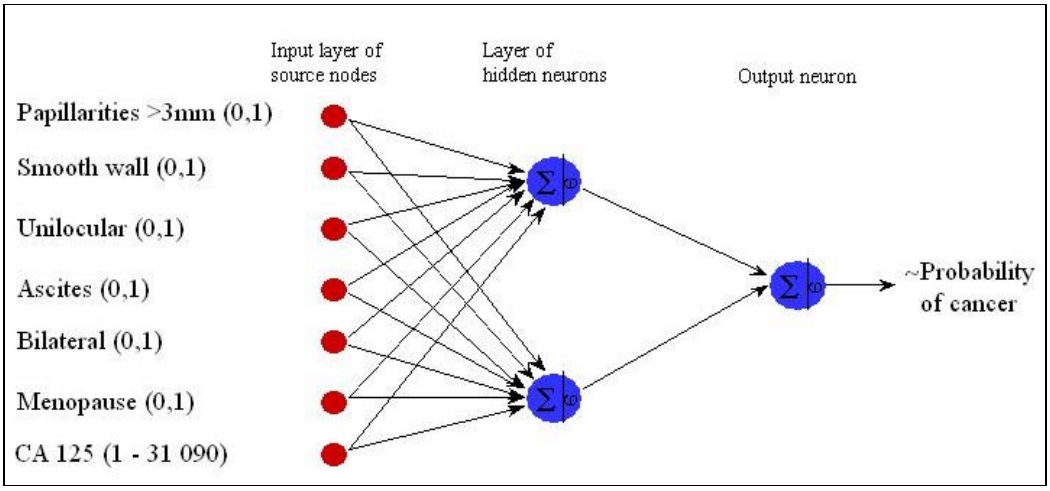


Figure 2.2. Structure of the best ANN that was used for prediction of malignancy of the tumors. This ANN incorporates more ultrasonographic data than the first ANN. (Timmerman et al., 2000)

Both ANNs are multilayer feedforward networks, containing one hidden layer. The left layer of nodes is the input layer, where values for the input variables enter the network. The middle layer is called the hidden layer, required for processing of the input values. The output layer returns a value for the outcome variable of interest. In the ovarian cancer case, we tried to predict malignancy of the tumor (output) based on age, menopausal status, serum CA125 level, and ultrasound data (inputs).

The ANNs are feedforward networks, which means that connections are only allowed from the input layer to the hidden layer and from the hidden layer to the output layer. Thus, each node in the hidden layer receives a value from each input node. This means that the value that is passed on from a hidden node to the output layer is based on all the input values. In basic ANNs, the hidden

nodes simply calculate a weighted sum of all input values. The output nodes perform a similar calculation based on the values they receive from the hidden nodes. In both layers the results of the summations are transformed using a nonlinear function before they are passed on to the next layer. The results of the weighted sums therefore represent the strength of the interactions in the network. The weights in the network are determined by training the ANN on a large set of observations from available data (training set), including both input values and the desired output. Then, the network tries to predict the correct output for each set of inputs by gradually reducing the error. This is done by changing the weights of a node by an amount proportional to the error at that node multiplied by the output of the node that is feeding into the weight. Training the network consists of two steps:

- Forward step:
  The output and the error at the output node are calculated.
- Backward step:
  The error at the output node is used to alter weights on the output node. Then, the error at the hidden nodes is calculated by backpropagating the error at the output node through the weights. Finally, the weights on the hidden nodes are altered using the latter values.

For each set of input values a forward step and backward step is performed. This is repeated over and over again until the error is small enough.

II.2.3. Bayesian networks

As said before, there are two different sources of information for predictive models: the biological and medical information available about the nature of the disease and the growing amount of patient data. Data-driven models such as logistic regression and ANN models do not exploit the prior knowledge available about the problem at hand. As a result, data-driven models are often extremely data hungry and thus need major data collection efforts. Knowledge-based models on the other hand, cannot make full use of the quantitative observations. Bayesian or Bayesian networks provide a solution to integrate efficiently the background knowledge and the observations. Bayesian networks have been successfully applied in a broad spectrum of applications in which the proportion of prior knowledge and of patient data varied widely.

Uncertainty is inherent to almost all medical problems. A compelling approach to managing various forms of uncertainty is to formalize the problem within a probabilistic framework. In particular, Bayesian statistics offers a solid theoretical foundation that makes it possible to express coherent subjective beliefs of human experts in a probabilistic way. Bayesian networks provide a practical tool to create and maintain such probabilistic knowledge bases. A Bayesian network is a knowledge model that can be used as the kernel in expert systems. Furthermore, Bayesian theory describes the integration of new observations to the probabilistic model. Consequently, Bayesian networks are a natural solution to integrate prior background knowledge and data.

A Bayesian network (see Figure 2.3) represents a joint probability distribution over a set of variables. The model consists of a qualitative part (a directed graph) and quantitative parts (dependency models). Directed graphical models are not allowed to have directed cycles and have a complicated notion of independence, which takes into account the directionality of the edges. For a particular domain, the vertices of the graph represent the domain variables and the directed edges describe the probabilistic dependency-independency relations among the variables in accordance with the joint probability distribution over the domain variables. There is a dependency model for every vertex (i.e., for the corresponding variable) to describe its probabilistic dependency on the parents (i.e., on the corresponding variables). These dependency models can be considered as input-output probabilistic models defined by a parametric family of distributions and a corresponding parameterization. If the variables are discrete, a common

dependency model is the table model, which contains the conditional distribution of the child variable conditioned on its parent.

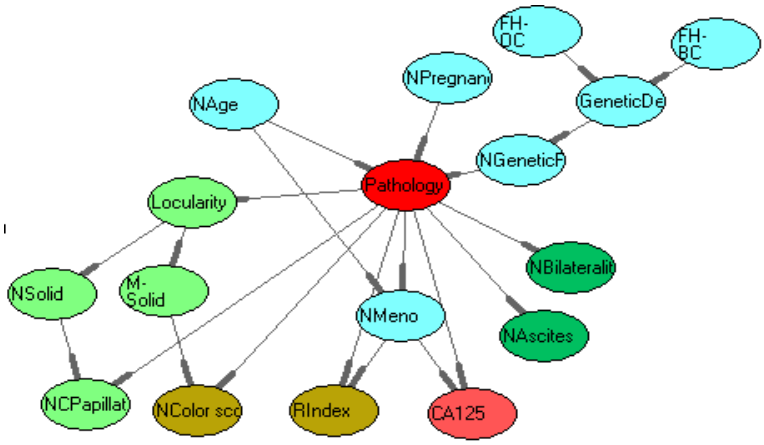


Figure 2.3. This Bayesian network represents the joint probability distribution of the measurements in the record of a patient with an ovarian tumor. Nodes represent the variables, such as age, pathology (benign vs. malignant), and CA125 serum level. Edges represent the probabilistic conditional dependency between variables. For example, the probability of the CA125 level being low, medium, or high given the menopausal status and pathology is independent of all other variables. Edges are quantified by a probabilistic model, such as a probability table. For example, the presence or absence of a genetic defect (GeneticD = 0 or 1) is a probability value for each of the configurations of the family history of ovarian cancer (FH-OC = 0 or 1) and family history of breast cancer (FH-BC = 0 or 1).

II.3. Design of the data

II.3.1. Endometrial cancer data

Data from 104 consecutive patients with endometrial cancer were prospectively collected. All patients were scheduled to undergo pre-operative ultrasound examination including Color Doppler Imaging. From this group, 97 women underwent full surgical staging. For these women, clinical and ultrasound data were collected. Figure 2.4 and Figure 2.5 show examples of sonographic and Color Doppler images of an endometrial adenocarcinoma.

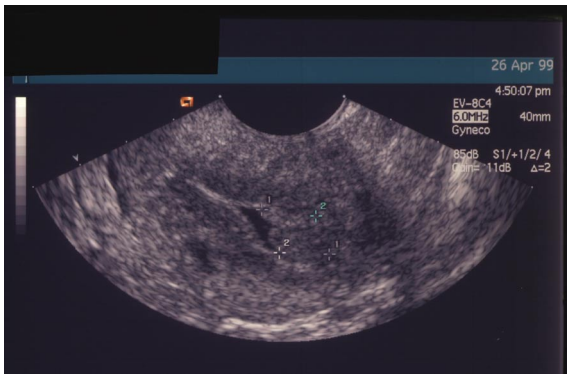


Figure 2.4. Transvaginal sonography (gray scale) of an endometrial adenocarcinoma.

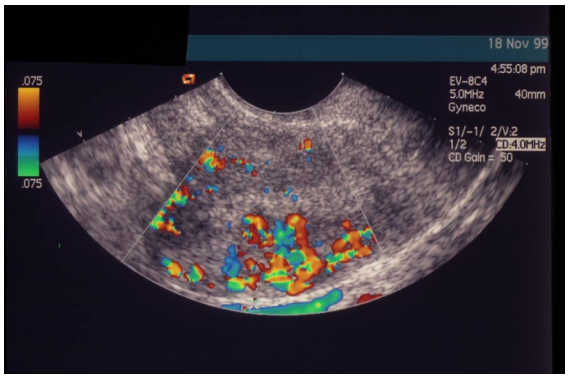


Figure 2.5. Color Doppler imaging of an endometrial adenocarcinoma.

II.3.2. Ovarian cancer data

Clinical data such as age, menopausal status and serum CA125 levels together with sonographic features of the adnexal mass were collected from 173 patients, scheduled to undergo surgical investigations at the University Hospitals in Leuven. The data originated from ovarian masses preoperatively examined with transvaginal ultrasonography. To compare this data from different groups all over the world, the IOTA consortium developed a specific protocol for the study. Figure 2.6 and 2.7 show examples of ultrasound images of adnexal masses in the ovary.

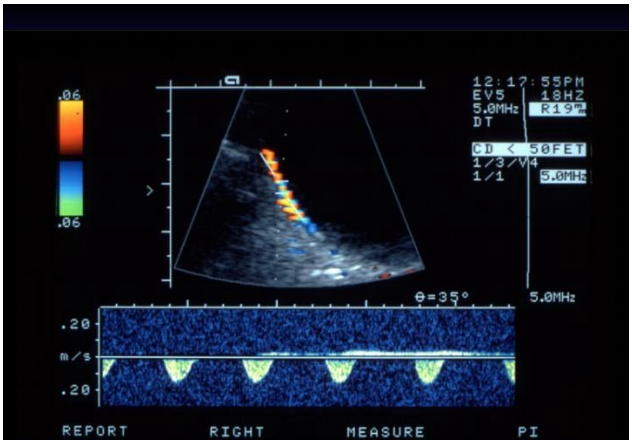


Figure 2.6. Ultrasound image of adnexal masses in the ovary.

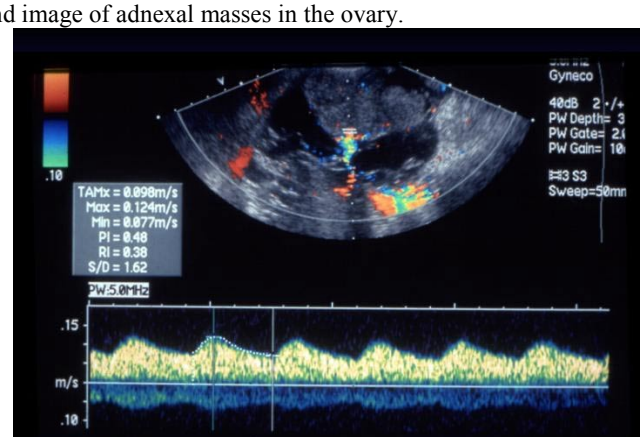


Figure 2.7. Ultrasound image of adnexal masses in the ovary.

Histopathological examination determined the presence of malignancy in each patient. The adnexal masses of 124 patients were found to be benign tumors while the remaining 49 patients were found to have a malignancy.

II.3.2.1. The IOTA protocol

For the collection of ultrasonography reports of patients with ovarian tumors, the IOTA protocol (Timmerman et al., 2000) describes extensively all patient variables (over 70) to be collected. Examples of issues addressed by the protocol are

- How should the variable be measured (e.g., the volume of the tumor is calculated from the three diameters in two perpendicular planes)
- Is the variable mandatory or optional?
- A controlled list of possibilities is given for variables with multiple options (e.g., the type of the tumor can be unilocular, unilocular-solid, multilocular, multilocular-solid, solid, or unclassified).

II.3.2.2. The IOTA database

A data model has been designed to store all patient variables in the format described in the IOTA protocol. The data model has been implemented using a Microsoft Access database that serves as the central repository for the patient records of all participating centers.

II.3.2.3. The IOTA web site

To provide a user-friendly and flexible way for the centers to enter their patient records, we have developed a web application using Active Server Pages and XML (eXtended Markup Language). This web site is available at http://www.iota-group.org and contains the following functionalities:

- Information pages for all visitors
  - Information about IOTA
  - A list of the participating research centers
  - Information for the press and general public
  - The protocol in html and pdf format
  - Information about the registration of a new user
- A secure application for the entry of patient records
  - Access to the application is restricted with a username/password
  - All traffic between the client (a web browser) and the server is encrypted using Secure Socket Layer technology.
  - Patient data can be entered and updated, new tumor masses and new patient visits can be added, and a patient report can be viewed or printed.
  - The entry of every variable is strongly validated against the protocol. This way the database only contains correct and complete patient records without any inconsistency.
- An administrator module providing functions to manage users and groups

Since the opening of the web application in November 2000, 70 complete records per month are being entered on average into the central database. By now, the total number of cases exceeds 1000. Figure 2.8 and 2.9 show screenshots of the IOTA web site.

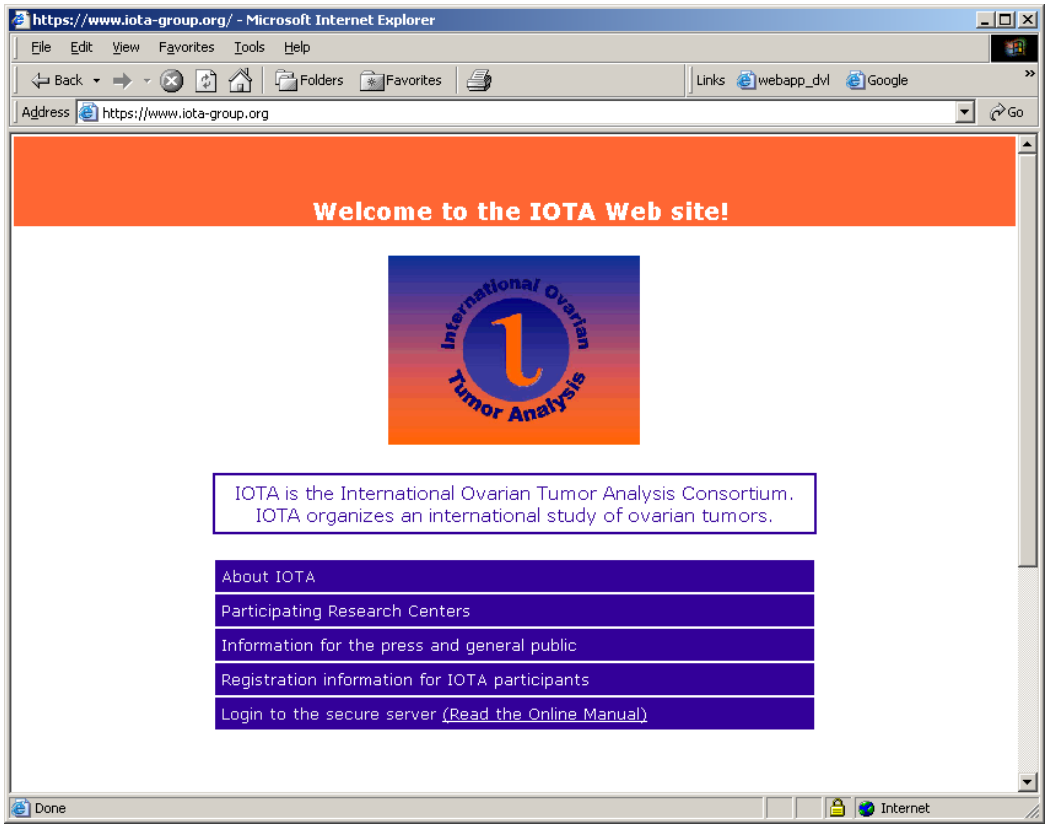


Figure 2.8. Homepage of the IOTA website at http://www.iota-group.org.

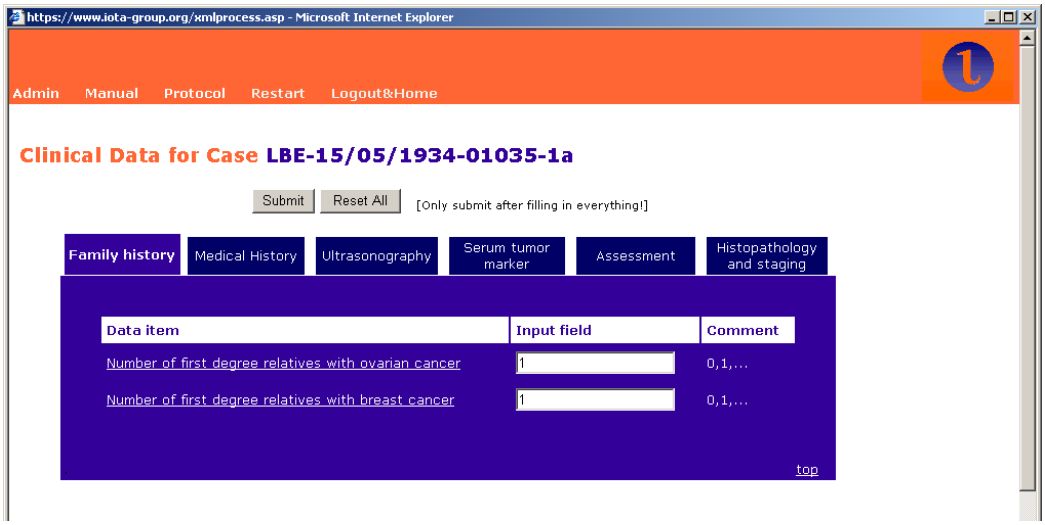


Figure 2.9. Data entry form, automatically generated from XML.

II.3.2.4. A general case reporting tool

Instead developing an application that is specific for the IOTA consortium, we have chosen a more general approach:

- All variables are defined in a XML file with their formats and value lists.
- All validation rules are also defined in a XML file.

- Based on these XML files, the web site and the validation code are generated automatically by the system.

This allows us to easily employ a case reporting tool that is similar to the IOTA application, but in another field of medicine with another protocol and other variables.
Moreover, because of the intended openness of the system, we were able to link the entry of a patient record with an immediate statistical feedback to the user (currently not available through the web site). Using an ANN trained on a set of 300 patient records; a probability for the malignancy of the ovarian tumor is given based on a subset of the variables.

II.4. Development of the models

II.4.1. Logistic regression for the prediction the depth of invasion in endometrial carcinomas

The logistic regression model aims to correlate clinical and ultrasound parameters with the depth of myometrial invasion. The myometrium is the muscle layer of the uterus. The depth of invasion of the tumor in the myometrium is an important prognostic factor and a key element in deciding the treatment schedule. According to the depth of myometrial invasion, endometrial carcinomas are separated into two classes, stage Ia and Ib (invasion below 50% of the normal myometrial thickness) and higher stage. Because of missing values, the coefficients of the logistic regression model were based on clinical and ultrasound (gray scale and CDI) data from 93 patients.

Multivariate logistic regression analysis was used for selection of the variables that are significantly correlated with the distinction of the two classes that are described above. Stepwise selection (logistic regression) retained four significant variables: the endometrial thickness (ET), the volume of the tumor (TV), the degree of differentiation (G1, G2) and the number of fibroids (NF). These four variables were included in the logistic regression model resulting in

$$y = -3.85 - 2.57\,NF + 2.49\,G1 + 2.61\,G2 + 0.21\,ET - 0.12\,TV$$

where G1 and G2 represent the degree of differentiation of the endometrial tumor. G1 is set to 1 if the carcinoma is moderately differentiated. In all other cases, G1 equals 0. Similarly, G2 equals 1 if the tumor is poorly differentiated. The logistic regression model gives an estimate of the probability p (see the subsection on logistic regression) that a given patient has deep myometrial invasion.

II.4.2. ANN models for the discrimination of malignant and benign adnexal masses

In this project, ANNs were used to predict the malignancy of the masses (Timmerman et al., 1999b). The collected clinical and ultrasound data were used as input variables in the ANNs to predict the outcome of the histological classification of excised tumor tissues in terms of benign or malignant (including borderline). The ANNs were trained on a randomly selected set of data from 116 patients and tested on the remaining data. For more information on the structure of the neural networks that were used in this project, we refer to Figure 2.1 and 2.2.

II.4.3. Bayesian networks for ovarian tumor diagnosis

One of the distinctive features of the discrimination between benign and malignant masses is the centrality of the status of the mass. Indeed, the data collection protocol excludes all other diseases and it ensures the presence of either a benign or a malignant mass. We use a single binary variable for discrimination. Taking advantage of the causality interpretation for Bayesian networks, this variable can “separate” the rest of the variables into two groups: causes (such as risk factors) and

effects (such as symptoms). We used the variables summarized in the left column of Table 2.1. The continuous and integer valued variables were discretized in accordance to the medical literature and expert knowledge.

Variable	Source of quantification
Family members with ovarian cancer	Expert
Family members with breast cancer	Expert
Genetic risk	Expert
Genetic deficiency	Literature, expert
Pregnancy	Expert
Age	Expert
Pathology	Literature, expert
Menopausal status	Expert,
Locularility	Literature
Color score	Literature
Resistance index	Literature
Bilaterality	Literature
Ascites	Expert
Papillation	Literature
CA125	Literature

Table 2.1. Sources of quantification for the *standard* model.

The building of the Bayesian model evolved through three different stages (Antal P. et al., 2001a). At the first stage, we experimented with "biological" models in which various causal models of the disease were incorporated. The specification of the structure was relatively easy, but the quantification was not possible from the literature, nor from the expert and we had too small a data set to quantify the many hidden variables that we add to introduce. At the second stage we built "expert" models that reflect the expert's experience. The specification of the structure of the Bayesian network was again relatively straightforward. However the results were too biased because the medical expert participating in the project previously worked with the same collected data, so his estimates were largely based on the data set. Finally, we decided to build "heterogeneous" models containing biological models of the underlying mechanism quantifiable by the literature (e.g., the genetic part), parts quantified by a medical expert (e.g., age, parity distribution of the patients), and parts quantified by previously published studies (such as the effect of locularity or blood flow). The final model, called *standard*, is shown in Figure 2.10 and a screenshot of the Bayesian network tool for prediction of malignancy of ovarian tumors is shown in Figure 2.11. The naïve models have no prior quantification (i.e., a priori specified dependency models for each variable).

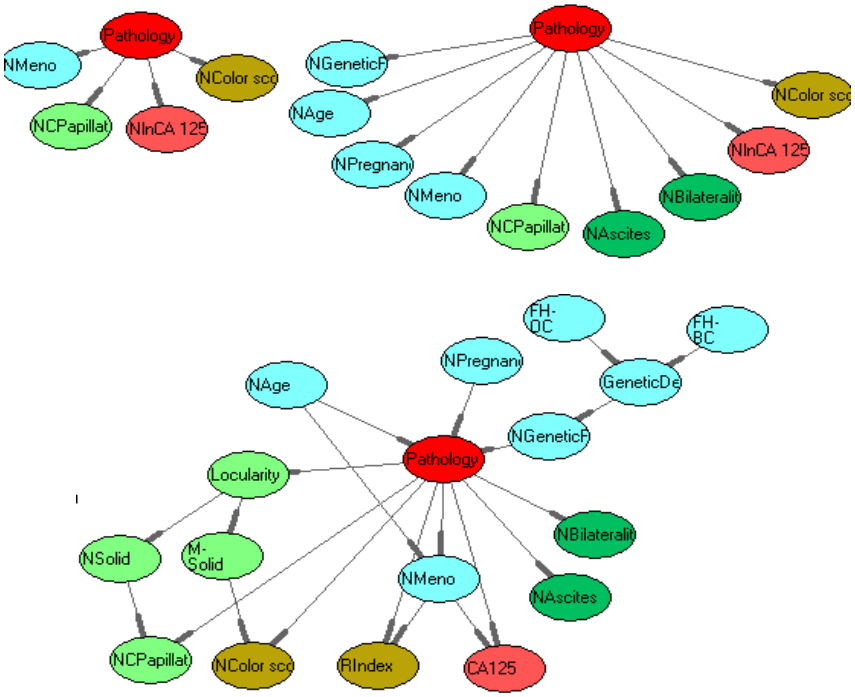


Figure 2.10. The *small naïve* (top left), *large naïve* (top right), and *standard* (bottom) Bayesian networks.

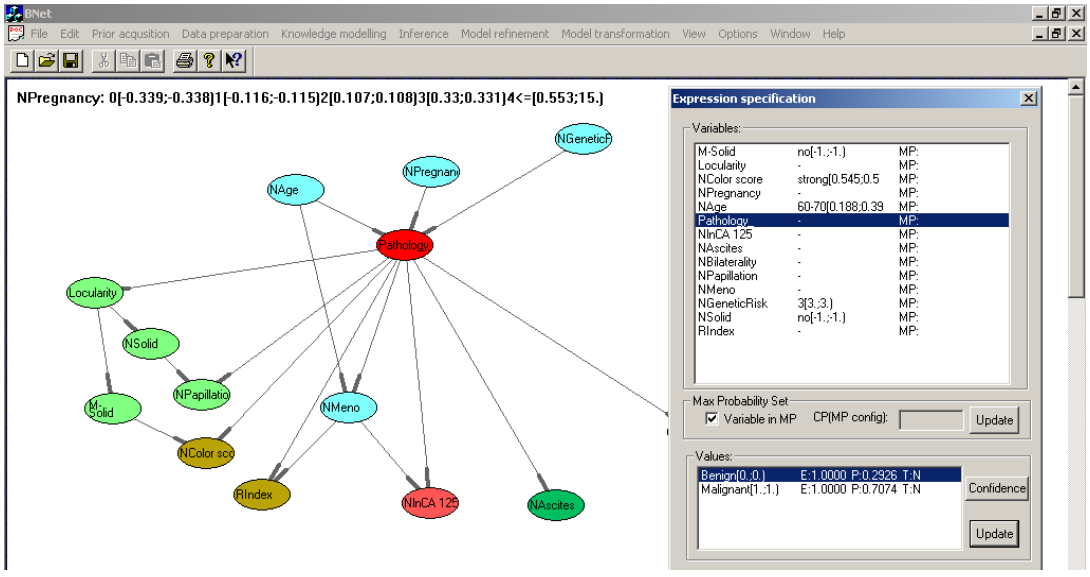


Figure 2.11. Screenshot of the Bayesian network tool that was developed for the diagnosis of ovarian tumor malignancy.

Because of the extensive and complex usage of the prior knowledge we used a strict documentation method to track the route of the prior information from studies into the model. Conversion formulas were constructed to compile the raw prior knowledge to be compatible with

the conditions of the task and the format of the Bayesian network. The following list contains the high-level steps of this process:

1. Make a list of all prior knowledge about variables, discretizations, existing dependency models, and so on.
2. Classify the different types of priors available (from exactly specified prior submodels to high-level guesses about qualitative dependencies).
3. Select a "coverable variable set" what seems to be quantifiable from the prior background knowledge and the available data.
4. Specify a complete domain model by following the standard construction mechanism for Bayesian networks and considering the existing prior submodels.
5. Construct secondary conversion models and formulas to quantify the final model (incorporating hyperparameters about confidence, conditioning on the conditions of the discrimination task, and so on).
6. Quantify the model based on the available data, documenting the sources of the information for interpretation, modification, and maintenance.

II.5. Evaluation of the models

II.5.1. Evaluation of the ANN model

The performance of each ANN model (Timmerman et al., 1999b) was evaluated using ROC (receiver operating characteristic) curves and compared with the corresponding results of the RMI model (Tingulstad et al., 1996), two logistic regression models (Tailor et al., 1997; Timmerman et al., 1999a) and a model exclusively based on morphological data (Lerner et al., 1994). ROC curves are applied in the medical world to measure the accuracy of a test in identifying diseased cases.

When the results of a particular test in two populations are considered, one with benign and the other with malignant tumors, a perfect separation between the two groups will be rarely observed. Indeed, the distribution of the test results will overlap, as shown in the Figure 2.12.

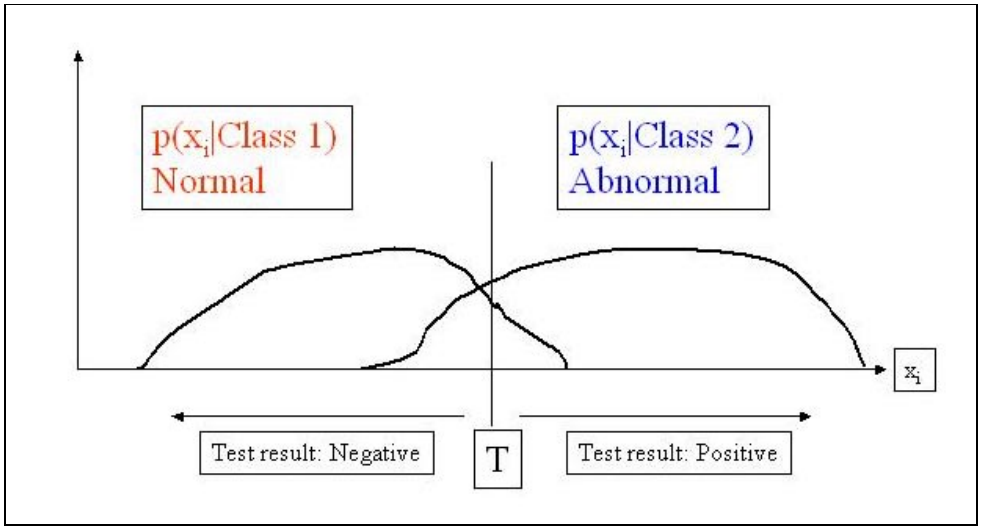


Figure 2.12. Overlap of the results of a test for classification of patients into two groups, normal (benign tumor) and abnormal (malignant tumor).

For every possible threshold that is selected to discriminate between the two populations, there will be malignant cases that are correctly classified as positive (true positives), but some malignant cases will be classified negative (false negatives). On the other hand, a number of

benign cases will be correctly classified as negative (true negatives), but some benign cases will be classified as positive (false positives).

In a ROC curve, the true positive rate is plotted as a function of the false positive rate. The true positive rate or sensitivity and the false positive rate (which is equal to one minus the specificity) are expressed as percentages of patients. The ROC curves for the classical tests are given in Figure 2.13.

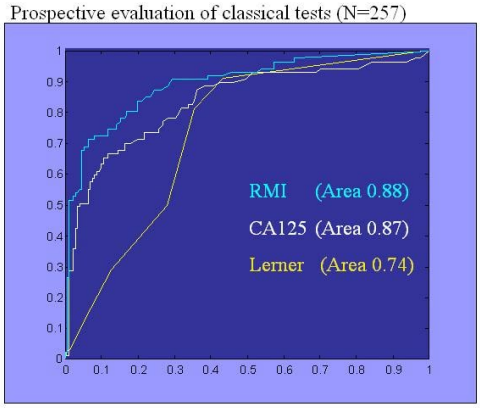


Figure 2.13. ROC curves of the Risk of Malignancy Index, a model exclusively based on morphologic data (Lerner et al., 1994) and serum CA125 levels.

The ROC curves for two logistic regression models and the best ANN are shown in Figure 2.14.

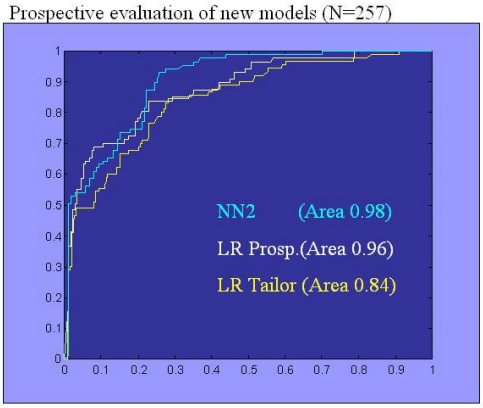


Figure 2.14. ROC curves for the best ANN and two logistic regression models (Tailor et al., 1997; Timmerman et al., 1999b).

The results of the ROC evaluation are given in Table 2.2.

	Area under ROC curve	**Sensitivity**	**Specificity**
Morphology (Lerner)	0.74	81	64
RMI	0.88	67	91
Log Regression	0.96	96	85
ANN2	0.98	96	93

Table 2.2. Results of the evaluation of the performance of the best ANN compared to the RMI and the best logistic regression model.

The specificity, sensitivity, and the area under the ROC curve of the best ANN model were 95.9 %, 93.5 %, and 0.979 respectively. The corresponding values for the RMI model were 67.3 %, 91.1 %, and 0.882 and for the logistic regression model 95.9 %, 85.5 %, and 0.956.

The specificity is the probability that the test result (RMI, result of logistic regression, outcome of ANN) will be negative when the tumor is benign (true negative rate). The value for the area under the ROC curve can be interpreted as follows: an area of 0.84, for example, means that a randomly selected individual from the group of patients with malignant tumors, has a test result larger than that for a randomly chosen individual from the group of patients with a benign tumor in 84% of the cases (Zweig and Campbell, 1993). In other words, the area under the ROC curve is a measure of the overlap between the test results of the two populations of patients (see Figure 2.12). When the test result under study cannot distinguish between the two groups, the ROC curve will coincide with the diagonal and the area will be equal to 0.5. When there is a perfect separation of the test results of the two groups, the area under the ROC curve equals 1 (the ROC curve will reach the upper left corner of the plot).

II.5.2. Evaluation of the logistic regression model

Using ROC curves, the performance of the logistic regression model was compared to the performance of important single variables in predicting the depth of myometrial invasion and the results of the subjective assessment by the gynecologist. The results of the ROC analysis are shown in Figure 2.15 and in Table 2.3.

Model / Variable	Threshold	Area under ROC	Sensitivity (%)	Specificity (%)	Accuracy (%)
Myometrial thickness	8 mm	0.706	74	61	66
Endometrial thickness (ET)	14 mm	0.762	81	64	71
ET / A-P diameter	0.429	0.754	72	71	72
Endometrial volume (EV)	4.93 mL	0.758	71	69	70
EV / uterine volume	0.085	0.775	69	80	76
Subjective assessment	-	0.787	61	86	76
Logistic regression model	0.28	0.904	85	81	83

Table 2.3. Performance of single variables, the subjective assessment of the gynecologist and the logistic regression model in the prediction of deep myometrial invasion.

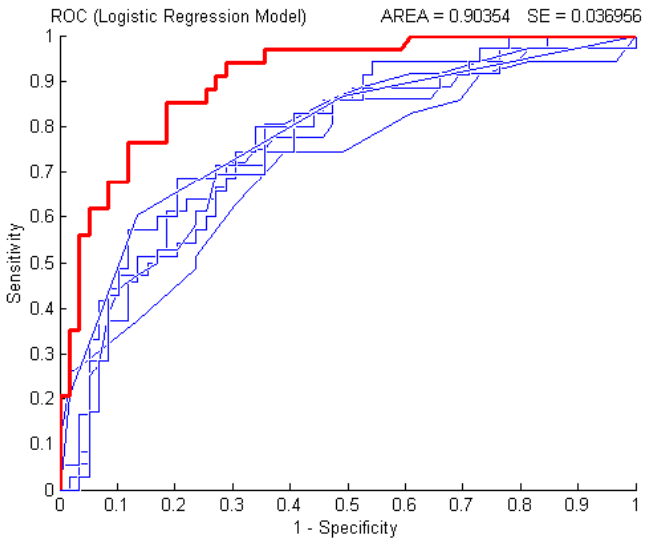


Figure 2.15. Prediction of deep myometrial invasion: ROC curves of the single variables and of the subjective assessment of the gynecologist (blue lines – also see Table 2.3) and of the logistic regression model (red line). Note that the logistic regression model performs significantly better (p=0.037, two-sided).

II.5.3. Evaluation of Bayesian networks

For comparison, we used the small and large naïve Bayesian network that assume complete conditional independence between the observations conditioned on the type of the mass. Five Bayesian network models are investigated: the two naïve models and the *standard* model in three contexts: prior quantification without hyperparameter update (i.e., without learning), update from the data without prior quantification and prior quantification with hyperparameter update. The performance is assessed with respect to the area under the ROC curve. Additionally, Table 2.4 contains the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The hyperparameters in the Bayesian networks are updated using 75% of the data set; the rest is used as a test set for estimating the area under the ROC curve (averaged over 1000 cross-validation sessions).

Model	ROC (%)	SE(%)	Sensitivity	Specificity	PPV	NPV
Serum CA125 (U/ml)	87.4	3.4	79.6	81.5	62.9	91.0
Risk of malignancy index	89.1	3.2	87.8	74.2	57.3	93.9
Logistic regression	90.4	6.0	85.7	81.1	63.2	93.8
Artificial neural network	95.1	3.9	87.5	92.7	82.4	95.0
Small naïve (BN)	93.1	3.9	94.7	74.1	84.2	90.0
Large naïve (BN)	93.8	3.7	96.6	79.9	90.1	92.6
Standard (prior + no update) (BN)	90.4	2.3	93.6	72.3	81.1	89.8
Standard (no prior + update) (BN)	95.0	3.4	94.2	83.5	84.8	93.8
Standard (prior + update) (BN)	95.2	3.4	94.7	83.4	86.1	93.7

Table 2.4. Performance of the models.

The risk of malignancy model is based on the menopausal status, CA125 serum test and on a morphologic score. The logistic regression model, the artificial neural network model and the small naïve Bayesian network have the same four inputs: menopausal status, CA125 serum test,

color score, and papillarities. The previously reported results were achieved on a smaller data set under different testing conditions. Comparison of the results in Table 2.4 considering this difference shows that the Bayesian networks have performances similar to that of the best artificial neural network. Bayesian networks are significantly better than the RMI and CA125. Although the standard prior Bayesian network without update has a slightly better performance than the RMI, the difference is not significant.

II.6. Discussion

Both for the ANN (Timmerman et al., 1999b) and the Bayesian network models (Antal P. et al., 2001a), the results show that these models can be used as a diagnostic for malignancy of adnexal masses, in combination with the opinion of the expert (gynecologist). Of all the tested models, both model were the most sensitive and performed significantly better than other models. Specifically for the Bayesian networks, our experience confirms that these models provide a practical solution for representing medical knowledge, performing inferences, and learning. They are particularly effective in integrating various prior submodels together. For example, the *standard* Bayesian network (quantified from previous studies about various subparts of this model) has the same performance as the RMI score, which is the accepted reference method. The negative effect of the crude discretization scheme is compensated by the multimodal approach in an appropriate dependency structure. Additionally, the prior quantification is dominated by a sample containing 50 to 100 random patient cases, which means that in our case (300 patients) the prior has no effect on the final performance. An important advantage of such white-box models is that they can be used for explanation or a semantic sensitivity analysis. Furthermore, because of their decomposed nature, they can be extended to perform finer subclassification, which is the next phase of our project. It can be concluded that the combination of expert assessment and the predictions of the ANN or the Bayesian network models will reduce the necessity of histological examinations. As a result, a significantly reduced number of patients will have to undergo the cumbersome experience of an operation.

For the endometrial carcinomas, it can be concluded that single morphological or color Doppler parameters cannot reliably predict the depth of myometrial invasion. Moreover, the latter parameters do not contribute in the logistic regression model since they were not retained after multivariate analysis.

II.7. Annotated Bayesian networks

Although we have obtained performance comparable to the human experts and to other statistical models, a specific limitation of Bayesian networks is that it requires a lot of work to document the prior knowledge we incorporate into the model. The amount of information from studies, papers, and experts linked to the Bayesian network is large. Consequently, we investigated methods to attach informal medical information to the formal probabilistic model and utilize this in the model-building phase and in decision support in general (Antal P. et al., 2001b).
To solve this problem at first we have defined a heuristic annotation method for Bayesian networks and modified our modeling environment accordingly. This annotation method is called the Annotated Bayesian Network and allows the attachment of natural language comments and references at various levels of a Bayesian network, such as model, groups, variables, values of the variables, the probabilistic dependency relation between the variables, and the corresponding probability distributions of the variables. In the next step, to fully exploit the potential of the probabilistic model enriched with textual information, we have defined a formal framework for this annotation by introducing an adjoined semantic network to which the textual information is attached. From the Bayesian network with its strict probabilistic semantics, we can derive a semantic network that describes explicitly the entities and their relations in the corresponding domain (i.e., domain ontology).

The most immediate application of such semantic domain model or ontology is the possibility to organize the textual information of a clinician (relevant papers, interesting patient records, important observations) by attaching it to various parts of the derived semantic network. An Annotated Bayesian Network can also provide a more detailed and even personalized explanation for the derived conclusion. The underlying probabilistic model can serve as a decision theoretic computational engine to select optimal decisions and to generate plausible explanations, while the corresponding semantic network with its natural language comments and references can enrich and personalize such explanations. Finally we have investigated the applicability of this model in an increasingly important task, namely in helping either knowledge engineers or clinicians in finding new textual information in large knowledge bases such as document or abstract collections or the Internet. We shortly describe methods that use the derived semantic network and the attached textual information to make the information retrieval more efficient. This is accomplished by expanding the query of the user, by influencing the relevance evaluation function and by ordering and/or filtering the resulting documents.

II.7.1. Semantic network representation of Annotated Bayesian Networks

In practical Bayesian networks, the theoretically complete representation of a probability distribution over a set of domain variables is augmented with

- Textual names for the domain variables and for the values of the discrete variables
- Discretization thresholds for the values of the discrete variables
- A partition of the variables grouping them in accordance with the domain semantics

Here we assume two further extensions: the existence of variable types (e.g., discretization schemes with value names) and the hierarchy of groups over the domain variables. The following description illustrates the representation of the extended Bayesian network shown on the left side in Figure 2.16:

1. Variables{values(bins)}: Pathology{benign, malignant}, CA125{normal(0,55), high[55,)}
2. Graph (directed edges): (Pathology,Papillation), (Pathology,Locularity), …
3. Dependency models: Pathology: Table{0.7,0.3,…}, …
4. Groups: Observation = {Morphologic, …}, Morphologic = {Papillation, Locularity}, …

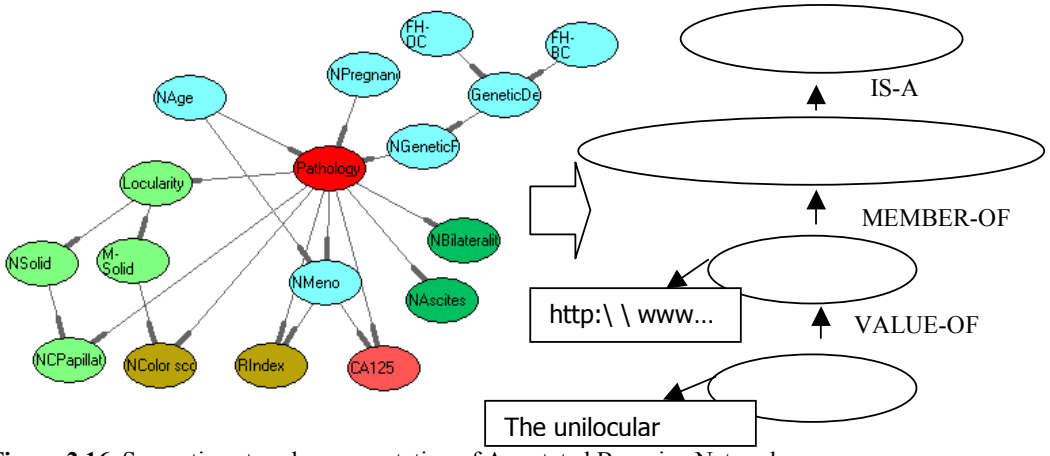


Figure 2.16. Semantic network representation of Annotated Bayesian Networks.

While such a representation is fully functional both for the knowledge engineers and the users of a Bayesian Network, we have noticed that the attachment of informal resources (such as natural language documents, comments, or figures) helps significantly the development and the maintenance of the model. Consequently, we further augmented our representation of Bayesian

networks by allowing attachment of a free text to the objects in the representation (i.e., values, variables, edges, dependency models and groups).

The Annotated Bayesian Network proved particularly useful for handling the heterogeneous sources of information in the Bayesian network used for ovarian tumor classification. The initial heuristic extension of the original representations has resulted in a hybrid representation: the Bayesian network representation with its strictly defined, but implicit probabilistic semantics (e.g., the (in)dependency model) and the augmentations with their domain specific, explicit semantics (e.g., the groups and their hierarchy).

Figure 2.16 illustrates the semantic network representation for Annotated Bayesian Networks and the list below describes the transformation of a Bayesian network and partly its semantics into the semantic network representation. Informally, the standard Bayesian network is incorporated automatically in the foundation of the domain ontology and the higher-level concepts are built on the top of it.

1. Value ⇒ Value node and *value-of* edge
2. Variable ⇒ Variable node and *is-a* edge
3. Group ⇒ Group node and *is-a* edge
4. Dependency model ⇒ Dependency model node and *input-of*, *output-of* edges
5. Annotation ⇒ Annotation node and reference edge

For simplicity, in the following the Annotated Bayesian Network denotes its semantic network representation.

II.7.2. Annotated Bayesian Networks in information management

The Annotated Bayesian Network incorporates partly the standard Bayesian network semantics, which already offers hierarchical and lateral relations between the concepts (e.g., variable-value and parent variable-child variable). The hierarchy of the groups, however, makes it possible to represent the ontology for a particular domain. Since the domain ontology represents the relevant concepts and relations in the domain, such representation is an ideal candidate for organizing the documents (i.e., electronic articles) of the knowledge engineer, who develops the model and ontology or of the user of the finished model and ontology. In the case of a knowledge engineer, the information sources used to develop the model can be attached to the corresponding objects in the model, while a researcher or a clinician can organize, store, and access his or her collection of relevant articles, patient records, and so on according to the domain semantics. In our experimental environment, we allow references to local files and documents on the network, and provide a browsing functionality for Annotated Bayesian Networks. In Figure 2.17, an example of an Annotated Bayesian Network is shown.

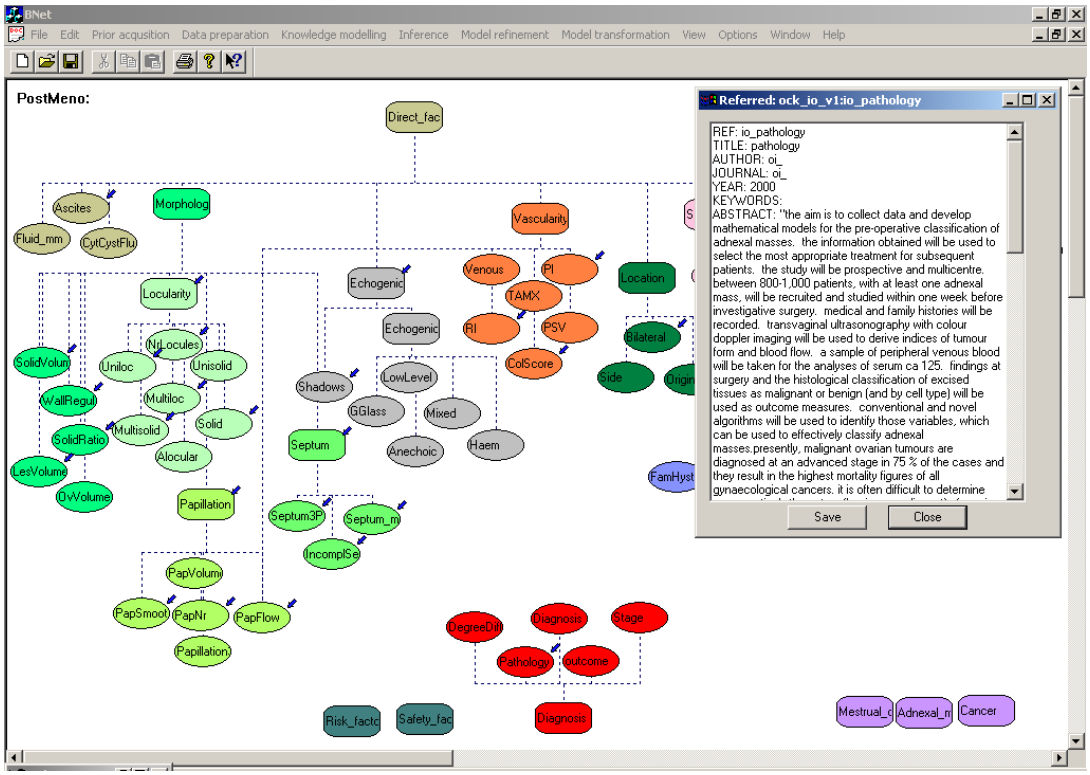


Figure 2.17. This Annotated Bayesian Network represents all the variables in the ovarian tumor domain together with their semantic relationships and the annotation for the variable *pathology*.

II.7.3. Annotated Bayesian Networks in decision support

In developing the Annotated Bayesian Network representation our central goal was to link a strictly defined computational domain model with a more flexible, mainly textual domain model. Our goal was on the one hand to help the manual development of the computational model as described above and on the other hand to help decision support. Bayesian networks or the related Influence Diagrams used in decision support are frequently used as input-output models, predicting the conditional distribution of an output variable conditioned on the observed values of the input variables. For example, in ovarian tumor classification, the Bayesian network is used to predict the conditional probability of malignancy given the observations about the morphology, the vascularization, and other properties of the mass. Although the performance of the model based on the predicted conditional class probability is comparable to that of human experts, an explanation of the computed prediction is frequently necessary. We assume that the predictive explanation in Bayesian networks consists of the most influential observed variables and the most influential paths connecting these variables to the output variable. Since this explanation set and the paths can be identified automatically, the attached annotations can be used either directly by offering them for browsing or indirectly by analyzing the attached textual information. The analysis can be a deep linguistic analysis to provide natural language excerpts or summaries for the variables involved in the explanation. Another approach, which we follow, uses a shallow statistical analysis. We experimented with generating vector representations for the annotated objects in the Annotated Bayesian Network, which contains relevance factors for each textual term occurring in the annotation. Based on the vector representation, the explanation derived from the computational model can be enriched, providing for example the average of the relevance factors for the variables in the explanation set or in certain path.

II.7.4. Information retrieval using Annotated Bayesian Networks

Information retrieval (IR) deals with methods for indexing, searching, and recalling data, particularly text and other unstructured types of data. Two major trends leading to increase in the efficiency of the IR process are to use user-specific information and domain-specific information. The Annotated Bayesian Network can serve both as the user profile and as the domain ontology. Figure 2.18 shows the standard model of the IR process (left) and a modified model based on Annotated Bayesian Network (right) with the following function support:

1. Annotated Bayesian Networks as domain ontology
    a. Deriving a domain specific controlled vocabulary from an Annotated Bayesian Network.
    b. Using the Annotated Bayesian Network as domain ontology for generalizing or specializing query terms.
2. Annotated Bayesian Network as user profile
    a. Expanding or transforming the query according to the Annotated Bayesian Network.
    b. Modifying the query matching formula to require that the relevant documents should be close to the aggregate Annotated Bayesian Network document (e.g., concatenation of all the annotation).
    c. Filtering the returned documents with respect to the Annotated Bayesian Network.

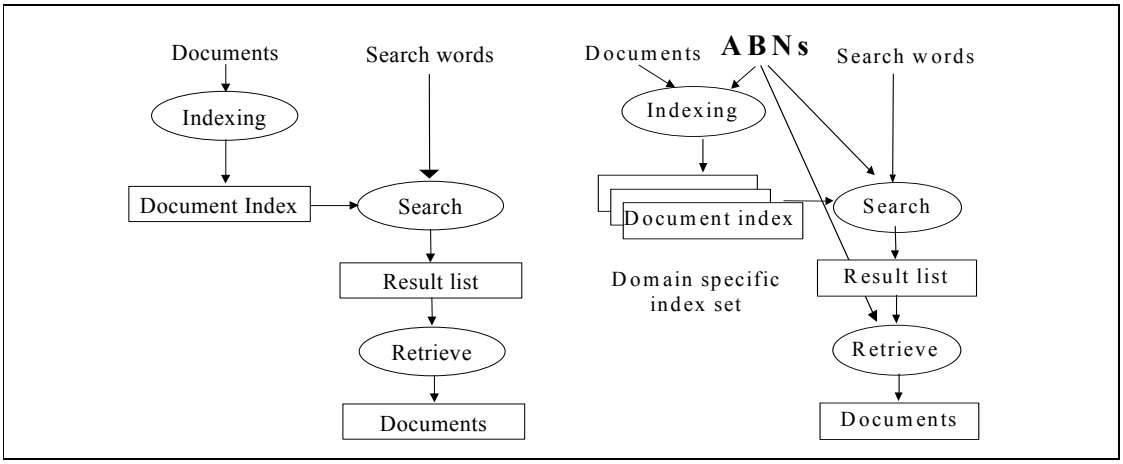


Figure 2.18. Information retrieval models: general (left) and based on Annotated Bayesian Networks (right).

Our main purpose was to increase the efficiency of the model building by helping the medical experts and knowledge engineers to find the necessary information. It means that we have attempted to integrate our Bayesian network modeling tool and an IR system. We created an integrated environment in which the computational and statistical modeling is accessible to the IR system as a kind of context (user profile) to increase the efficiency of the retrieval. The core of this system is the ABN. The availability of a computational-statistical domain model with textual annotations provides the link to integrate the both types of functions (e.g., probabilistic reasoning and learning versus information retrieval). In the IR subsystem the documents are represented by a vector representation. To improve the performance of the system we used a controlled index to represent the documents, in which the domain vocabulary was derived from the Annotated Bayesian Network.

Annotated Bayesian Networks are use to characterize the interest of the knowledge engineer more precisely. To this end, we identified phases in the model building, such as finding new domain variables, defining variables, checking conditional pair-wise (in)dependencies, quantifying dependency models, and so on. The query is automatically modified according to the actual model-building phase, so the document matching criteria is always biased towards the actual goal of the model building process. Finally, the returned documents are filtered to eliminate the redundancy.
The integrated system consists of two components: the statistical-computational modeling system to work with Annotated Bayesian Networks and a client-server document system for storage, analysis, and retrieval.

II.8. Conclusion

We have presented several mathematical approaches to medical decision support: basic statistical measures, logistic regression, neural networks, and Bayesian networks. Two specific diagnostic problems have been tackled: the preoperative prediction of malignancy in endometrial carcinoma using ultrasonographic measurements and the preoperative assessment of adnexal masses (ovarian tumors) using ultrasonographic measurements, clinical variables, and the CA125 serum marker. The logistic regression, neural networks, and Bayesian networks for ovarian tumors deliver performance comparable to that of a skilled gynecologist but lower than that of our expert. But for endometrial carcinoma (for which the expertise is much less extensive), the performance of logistic regression is at least as good as that of the expert. These performances demonstrate that machine learning techniques have a real added value for medical decision support in practice.

We also presented several software tools. First, a general case reporting tools support the data collection for the International Ovarian Tumor Analysis consortium. This web application has served for the collection and validation of the largest database of case reports on ultrasonographic assessment of patients with ovarian tumors. Second, an environment for Bayesian networks allows the development of knowledge-modeling and decision-support models.

Additionally, we have presented an extension of Bayesian networks called Annotated Bayesian Networks that let knowledge engineers keep track of the large volumes of documentation necessary for building complex Bayesian networks. Beyond information management, this methodology is also efficient in decision support and information retrieval. Annotated Bayesian Networks are part of our software environment for Bayesian networks.

III. Microarrays in oncology: from disease management to fundamental research

Genetic studies and molecular biology studies serve to elucidate mechanisms responsible for carcinogenesis. In the last few years, technological advances started a move from labor-intensive workbench studies (which produce only fragmentary results) to highly automated processes that collect massive amounts of data about the genetic basis of cancer and the behavior of tumors. Facing a deluge of information, biologists and physicians have turned to data mining to help assemble the millions of pieces of the puzzle that these new technologies churn out.

One of the most promising technologies recently developed in molecular biology is microarray technology (DeRisi et al., 1997; Lander, 1999; Schena et al., 1995). Microarrays make it possible to measure in parallel the activity or expression (transcription or amount of mRNA produced for a specific gene) of thousands of genes in a biological sample. From the biological point of view, the key idea is that by performing many microarray experiments, it is possible to collect enough clues to elucidate and integrate many molecular mechanisms that contribute to the development of cancer. From the point of view of information technology, the major challenge is that we must perform complex statistical analyses on data vectors of thousands of measurements (high-dimensional data is notoriously difficult to analyze).

Microarrays (which we discuss in more details in the next sections) have revealed themselves as an important tool for disease management (Alizadeh et al., 2000). Indeed, the global expression profile as measured by a microarray of a patient suffering from a specific disease can be used as a molecular fingerprint for that disease. The identification of such disease-specific fingerprints (e.g., expression profiles of tumor cell lines) can help in ameliorating the diagnosis, staging and prognosis in a way similar as described for the more classical clinical tests previously outlined.

However, microarrays go much further than disease management only. They reach their full impact in the study of fundamental biological processes. Because of their high-throughput nature, they have become a method of choice for the study of gene function and regulation on a global genomic scale.
Through ‘guilt-by-association’, genes of unknown functions can be assigned a putative function by linking them to genes with similar patterns of expression whose function is already known. In many organisms, genes for which nothing is known about the function still represent 30% of the genome and for many more genes the information available is fragmentary at best. Microarrays therefore provide a powerful approach to this extremely pressing question. Further, they are invaluable for unraveling the networks of regulation that control the dynamic behavior of genes. Understanding the network of interaction between genes is the central goal of genomics. Besides its intellectual value in its own rights and its many applications in biotechnology, this understanding lies at the foundation of an analytical approach to the management of pathologies and to drug discovery. Indeed, once the biological processes underlying the pathology are described in sufficient details at the molecular level, it becomes much easier to identify true drug targets. This fundamental approach has proved, time and again, its value in drug discovery. Needless to say, such studies of biological processes build a strong basis for further applications of microarrays in disease management. For these reasons and because microarrays are currently more actively used by molecular biologists than by clinicians, our work on microarray data analysis leans strongly towards the type of fundamental research just described.

The next sections and chapters strongly reflect this philosophy. First, we introduce the basics of microarray technology and discuss several issues around data storage and preprocessing. Second, we show how to use microarray data in building models for disease management in oncology. Third, we then switch to the analysis of microarray data for elucidating genomic processes. We

present an integrated methodology that combines microarray and genome sequence data to discover which patterns control gene expression at the sequence level. The first part of the methodology is the identification of groups of genes that show similar behavior. The underlying assumption is that, among the genes showing the same behavior (we say that they are co-expressed), some may share the same control pattern at the sequence level (we say that they are then coregulated). The second part of the methodology is the identification in such groups of which short motifs in the genomic sequence are likely to control gene expression. Fourth, we introduce a cutting-edge methodology that uses Bayesian networks to infer genetic networks of interaction from microarray data and prior biological knowledge and describe a general framework that integrates all the methods we present and many more into a knowledge pipeline for molecular biology.

III.1. Microarrays

Since microarray measurements and analysis are at the basis of the remainder of the chapters, this section will give a detailed description of the technology and the setup of a microarray experiment (see Figure 3.1 for an illustration of the robotic setup for making microarrays).

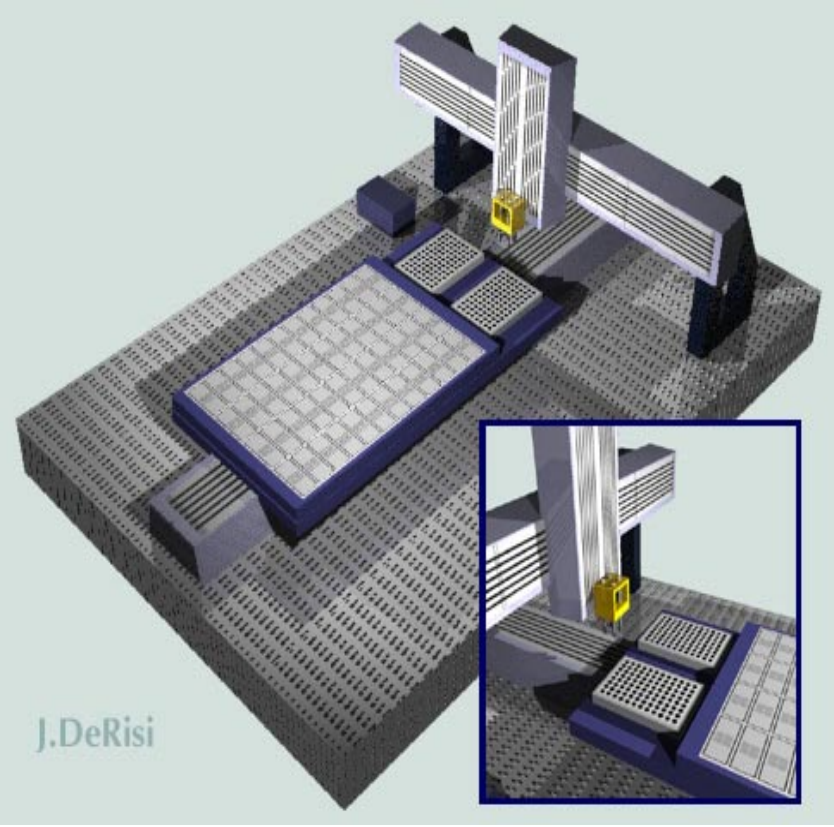


Figure 3.1. Microarray spotter.

Microarrays make it possible to measure in parallel the activity or expression (transcription or amount of mRNA produced for a specific gene) of thousands of genes in a biological sample (DeRisi et al., 1997; Lander, 1999; Schena et al., 1995). From the biological point of view, the key idea is that by performing many microarray experiments, it is possible to collect enough clues to elucidate and integrate many molecular mechanisms that contribute to the development of cancer. From the point of view of information technology, the major challenge is that we must

perform complex statistical analyses on data vectors of thousands of measurements (high-dimensional data is notoriously difficult to analyze).

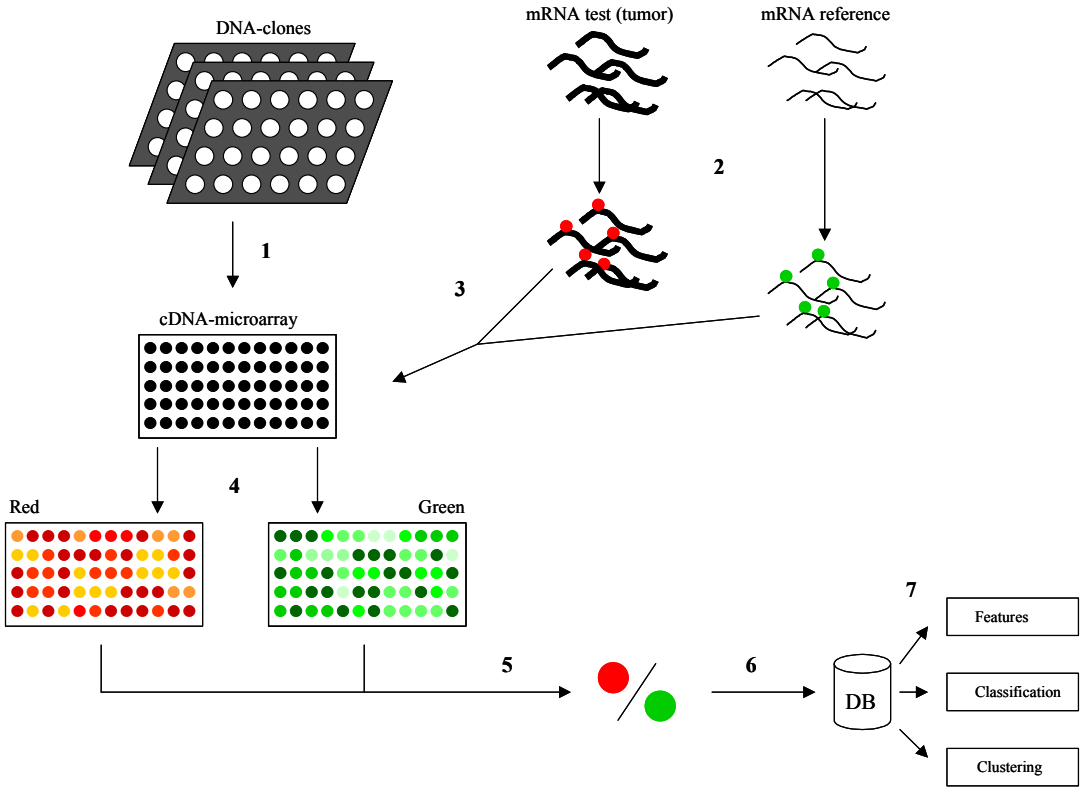


Figure 3.2. Schematic overview of an experiment with a cDNA-microarray. (1) Spotting of the pre-synthesized DNA-probes (derived from the genes to be studied) on the glass slide. These probes are the purified products from PCR-amplification of the associated DNA-clones. (2) Labeling (via reverse transcriptase) of the total mRNA of the test sample (tumor - red) and reference sample (green). (3) Pooling of the two samples and hybridization (4) Read-out of the red and green intensities separately (measure for the hybridization by the test and reference sample) in each probe. (5) Calculation of the relative expression levels (intensity in the red channel / intensity in the green channel). (6) Storage of results in a database. (7) Data mining.

Microarray technology allows simultaneous measurement of expression levels of thousands of genes in a single hybridization assay. An array consists of a reproducible pattern of different DNAs (primarily PCR products or oligonucleotides – also called probes) attached to a solid support. Fluorescently labeled cDNA, prepared from mRNA, is hybridized to the complementary DNA present on the array. cDNA-DNA hybridization intensities are measured by a laser scanner and converted to a quantitative read out. This data can be further analyzed by data-mining techniques (see below). Two basic types of arrays are currently available:

- Spotted arrays (Duggan et al., 1999) or cDNA-microarrays are small glass slides on which pre-synthesized single stranded DNA or double-stranded DNA is spotted. These DNA fragments are usually several hundred base pairs in length and are derived from ESTs (Expressed Sequence Tag) or known coding sequences from the organism studied. Usually each spot represents one single ORF (Open Reading Frame) or gene. A pair of cDNA samples is independently copied from the corresponding mRNA populations (usually derived from a reference and a test sample (e.g., normal versus malignant tissue)) with reverse transcriptase and labeled using distinct fluorochromes (green and red). These cDNA samples are subsequently pooled and hybridized to the array. Relative amounts of

a particular gene transcript in the two samples are determined by measuring the signal intensities detected for both fluorochromes and calculating the ratios (here, only relative expression levels are obtained). A cDNA microarray is therefore a differential technique, which intrinsically normalizes for noise and background. An overview of the procedure that can be followed with spotted arrays is given in Figure 3.2.

- GeneChip oligonucleotide arrays (Affymetrix, Inc., Santa Clara, CA) (Lipshutz et al., 1999) are high-density arrays of oligonucleotides synthesized *in situ* using light-directed chemistry consisting of thousands different oligomer probes (25-mers). Each gene is represented by 15-20 different oligonucleotides, serving as unique sequence-specific detectors. In addition mismatch control oligonucleotides (identical to the perfect match probes except for a single base-pair mismatch) are added. These control probes allow estimation of cross-hybridization. With this technology, absolute expression levels are obtained (no ratios).

Both techniques have their own advantages. Affymetrix technology claims to be more sensitive and allows absolute measurements. This is mainly because, for each gene, more probes are present on the array (including the mismatch probes). The major drawback of the Affymetrix technology is the relative high cost. cDNA technology is more flexible. Spotted arrays can be produced from both known and unknown cDNA and PCR fragments while for Affymetrix technology gene sequence information needs to be fully specified. Recent developments tend to use more cDNA probes per gene as to obtain more accurate performances.

III.2. Necessity for a laboratory information management system

As shown in the previous section, the production of a microarray is a complex procedure that is inevitably error prone. This necessitates the backtracking on several experimental settings or hypotheses. A recent study for example report error rates over 35% of the data points due to lack of consistency checks and flaws in annotation (Knight, 2001). Although it is impossible to render a system foolproof, guaranteeing an acceptable quality level and reproducibility is possible through meticulous recording of the various steps in the data generation process. In this section we will illustrate these issues and show how to automatically follow up and record the process of microarray production by using a Laboratory Information Management Systems (LIMS). The use of standards to this end can enhance and prolong the life cycle of a microarray experiment.

The following steps involved in the production of a cDNA microarray illustrate the many points, which are noise-sensitive or error-prone.

- Purchase or generation of proprietary ‘clone’ library – a collection of genetic fragments ideally representing a set of genes of a given organism. These clones are multiplied, stored and ordered into so called well-plates, usually per 96 or 384.
- Contamination (leaking of DNA fragments to neighboring wells) and equipment constraints usually require a reordering of the clones usually performed by pipetting robots.
- In the final preparation another robot ‘prints’ the genetic material to the glass slide so that each spot represents a single open reading frame or a gene. The biochemistry of this printing (or ‘spotting’) is complex and the possible configurations are numerous.

These procedures lead to a spotted glass slide or microarray, which can be used to conduct an experiment. To successfully backtrack on any errors that might have occurred during this labor-intensive production process, an automated administration of each action on the workbench is necessary.) LIMS serve this goal and we developed a first version for the VIB Microarray Facility to fully appreciate the nature of the data we are processing. Figure 3.3 demonstrates two screenshots of the GUI of this system. Behind are scripts that interact with the user and the database (cfr. direction of the arrows) to query, store or fetch the necessary information.

The hybridization(s) of test and reference sample (cfr. section using microarrays for oncology) on one or several microarrays and the consecutive measurement of the relative abundance of the screened gene transcripts represent the core of a single microarray experiment. Therefore it is connected in a modular way to the LIMS information; in Figure 3.4 we depict a structured database scheme that captures all the steps described above.

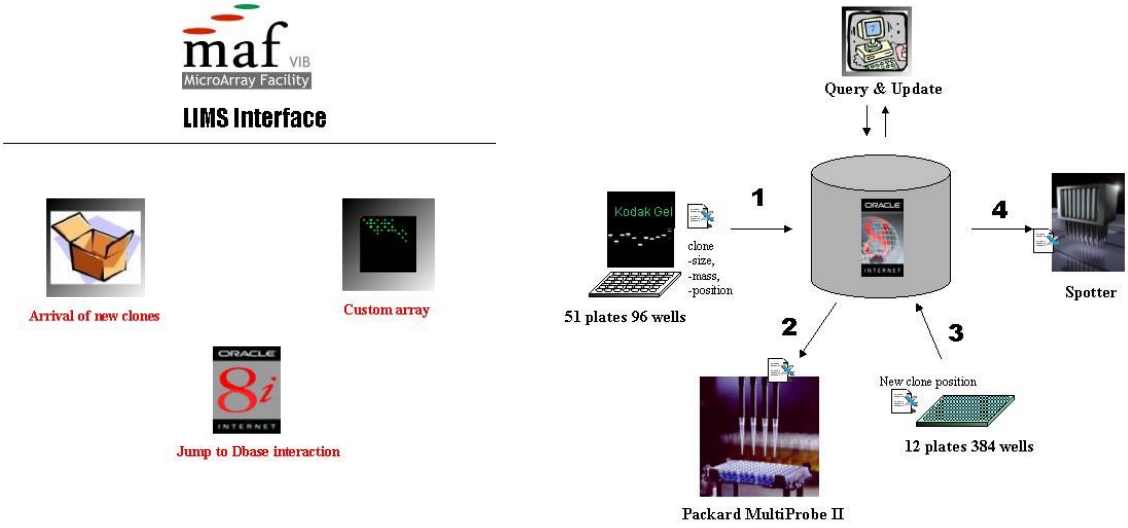


Figure 3.3. Screenshots of our Laboratory Information Management System: the various steps involved in a hybridization experiment are visualized for optimal tracking.

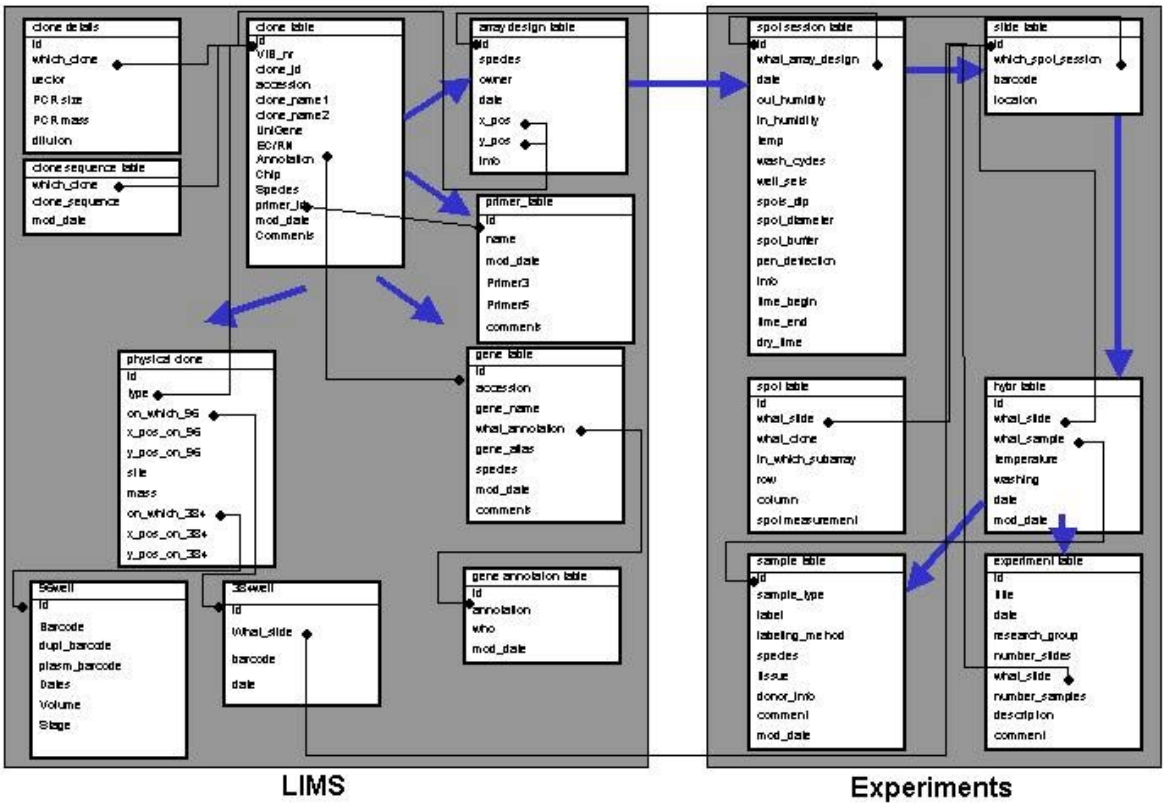


Figure 3.4. Custom gene expression database scheme.

III.3. Preprocessing of simple experiment design (black & white)

In the following, we explain the distinct sources of variation that consistently influence microarray measurements and hinder direct comparison of replicate experiments not measured under exactly the same conditions (i.e., measured on the same array, labeled with the same dye, and so on) (Schuchhardt et al., 2000 Yue et al., 2001). We show that microarray experiments are inherently noisy and that preprocessing methods can remove some of the variability caused by the experimental setup. Preprocessing methods derive for each gene the intrinsic expression level as caused by the condition tested. Many methods to preprocess the data have been used throughout literature. The choice of the method can have a major impact on the final results obtained. A thorough understanding of how different preprocessing methods transform the data and the search for a set of preprocessing techniques that make the data compatible with further analysis is a crucial aspect of microarray analysis. In this section, a standardized preprocessing flow will be proposed.

III.3.1. Sources of noise

To understand the necessity and importance of preprocessing, we need to have a clear picture of what raw data generated by a microarray experiment looks like. As a detailed example, we describe a simple black-white experiment based on a Latin square design (Figure 3.5). In such design, expression in two distinct conditions is compared (test and reference condition). On the first array, the test sample is labeled with Cy5 (red dye) while the corresponding reference is labeled with Cy3 (green dye). For each gene, two replicate spots are available on each array (referred to as left and right spot). In addition a color-flip experiment is performed: the same test

and reference conditions are measured once more in duplicate on a different array where dyes have been swapped. Such a design results in four measurements per gene for each condition tested, though not directly comparable with each other.

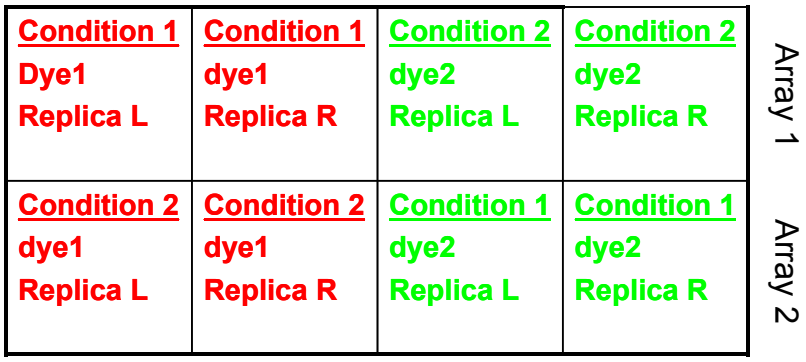


Figure 3.5. Schematic representation of a Latin Square design

A first set of effects that prohibits direct comparison between measurements are the "condition and dye effects". These effects reflect differences in mRNA isolation and labeling efficiency respectively between samples of the conditions tested. For genes equally expressed in both the reference and the induced sample the ratio of test/ref is expected to be 1. Condition and dye effects result in a deviation of such ratios from 1. The mathematical transformation that tries to compensate for these effects is called normalization.

A second source of variation is related to the imperfections of the spotting device used to produce the array. Small variations in pin geometry, target volume and target fixation cause spot dependent variations in the amount of cDNA present on the array. The observed signal intensity does not only reflect differences in mRNA population present in the sample but also the amount of spotted cDNA. Depending on the influence of the spot effects, direct comparison of the absolute expression levels might be unreliable. This problem can be alleviated by comparison of the relative expression levels (ratio of the test and reference intensities) instead of the absolute levels (see preprocessing). Indeed reference and test have been measured on the same spot and by dividing the measured intensities, spot effects drop out.

When performing multiple experiments (i.e., more arrays), arrays are not necessarily treated simultaneously. Differences in hybridization efficiency can result in global differences in intensities between slides making measurements derived from different slides mutually incomparable. This effect is generally called the array effect. All these effects occur simultaneously and prohibit direct comparison of expression levels. In Figure 3.6, two distinct preprocessing flows are depicted that are currently applied to our data prior to further downstream analysis (clustering, statistical testing, and so on). The different steps of these approaches will be explained in more detail in the following.

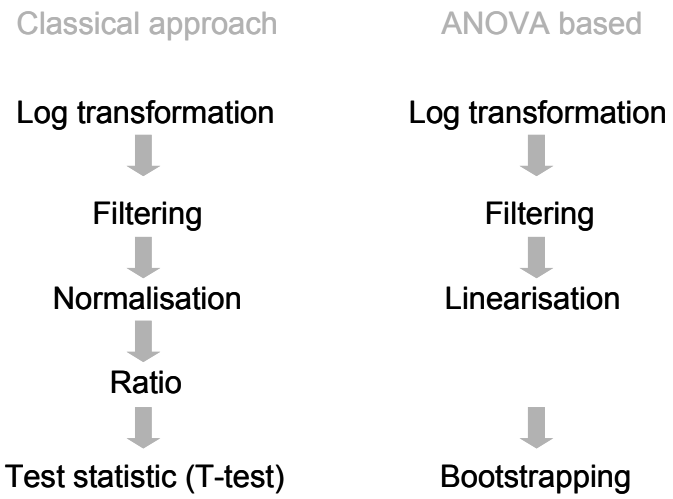


Figure 3.6. Flowchart of the distinct approaches described

III.3.2. Mathematical transformation of the raw data: need for a log transformation

A log transformation of the data is the initial step in the preprocessing data analysis flow. The necessity of this transformation is clear from Figure 3.7. In Figure 3.7A replicate measurements of different genes are plotted against each other. When considering untransformed raw data (with raw data we refer to background corrected intensity values), the increase of the residuals with increasing signal intensities clearly reflects the multiplicative effects. It is well known in signal processing that multiplicative errors decreasing the efficiency of most statistical tests (see further the underlying assumption of ANOVA models). It is therefore important to get rid of multiplicative errors by transforming the data.

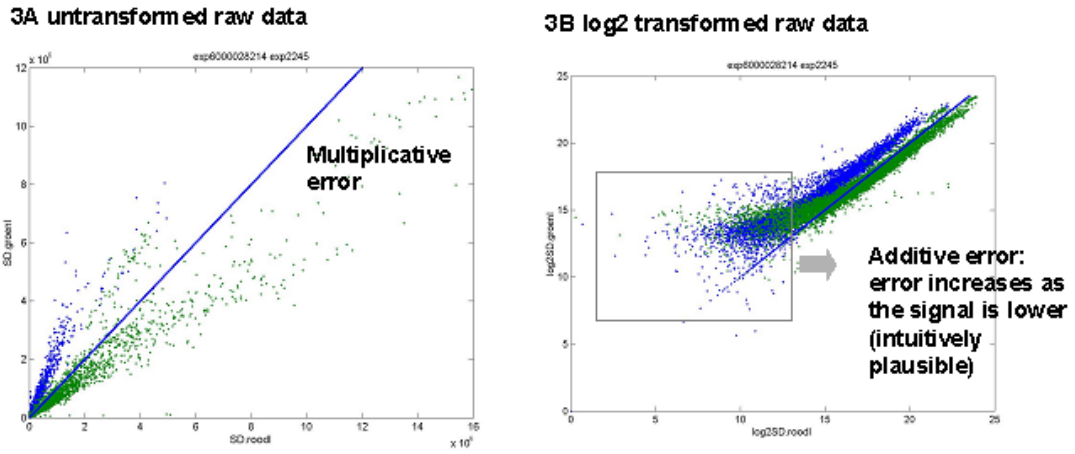


Figure 3.7. Illustration of the influence of log2 transformation on the multiplicative and additive errors. (A) representation of the untransformed raw data of both experiments (normal and color flip). *x* axis: intensity measured in the red channel, *y* axis: intensity measured in the green channel. (B) representation of the log2 transformed raw data of both experiments (normal and color flip). *x* axis: intensity measured in the red channel (log2 value), *y* axis: intensity measured in the green channel (log2 value).

After a logarithmic transformation multiplicative effects are considerably reduced as shown in Figure 3.7B. Indeed residuals are constant over a long range of signal intensities, except at low intensities where additive effects have been enlarged by the transformation. The increase of the

measurement error with increasing signal intensities as present in the untransformed data is counterintuitive since high expression levels are generally considered more reliable than low levels. In most cases log transformation of raw data is advisable (Baldi and Long, 2001, Kerr et al., 2000).

III.3.3. Filtering data

As a next step in the preprocessing flow, filtering is used to remove unreliable measurements or zero values from the data set. Such filtering procedures often depend on the choice of an arbitrary threshold (e.g., all genes of which the measured expression value does not exceed twice the expression level of the background are discarded). Since in our data sets the green and red channel display different sensitivities in the low expression level range, the choice of such threshold is prone to mistakes. Therefore, if possible we try to avoid filtering procedures based on an arbitrary threshold and try to retain all genes for further analysis (Kadota et al., 2001). The use of robust statistical tests allows the discrimination between statistically over- and underexpressed genes at a later stage of the analysis.

Zero values result in undefined values (e.g., when dividing by zero values or taking the log of a zero value) and therefore are automatically discarded for further analysis. However, in the light of a black & white experiment, zero values might be of major biological significance. Indeed consistent zero values correspond to genes switched off in one condition but on in the other condition. Therefore if a least one measurement for a gene contains a zero value in a particular condition, all measurements of that gene are treated separately.

We present two different approaches, known as the ratio technique and ANOVA technique, to perform normalization and to detect differentially expressed genes. The ratio technique is considered as the more classical approach for cDNA microarray analysis. More recently ANOVA techniques have been applied to microarray data.

III.3.4. Ratio approach

Normalization is a mandatory step to remove consistent condition and dye effects (see above). Although the use of spikes (control spots, external control) and housekeeping genes (genes expected not to alter their expression level under the conditions tested) has been described, global normalization is considered as most reliable. Global normalization assumes that only a small fraction of the total number of genes on the array alters its expression level and that symmetry exists in the number of genes that is upregulated versus downregulated. Under this assumption the average intensity of the Cy3 channel should be equal to the average Cy5 channel.

Based on the hypothesis of global normalization, the average ratio log2 (red/green) should equal 0. Regardless of the procedure used, all normalized log-ratios therefore will be centered on zero. In the following different procedures to perform global normalization will be described.

Linear normalization assumes a linear relationship between red (R) and green (G) intensities. A common choice for $c = \log_2 k$ is the mean or the median of the log intensity ratios for a given gene set. Alternatively the constant normalization factor can be determined by linear regression of the red signal versus the green signal. The coefficient as identified by regression determines the rescaling factor that should be used to either divide or multiply the red signal to obtain an average signal of 0 (in log scale).

As can be derived from Figure 3.8, the assumption of a constant rescaling factor for all intensities is an oversimplification. Indeed dye and condition effects seem to be dependent on the measured intensity. Such intensity dependent patterns are better visualized using a plot of M versus A (see Figure 3.8 for definitions of M and A). The relationship between the dyes behaves linearly only in

a certain range. However, when measured intensities are extreme (either high or low) nonlinear effects occur. To take into account these nonlinear effects during normalization, we prefer to use a robust scatter plot smoother Lowess that performs locally linear fits and allows calculating the normalized absolute values log2(*R*) and log2(*G*).

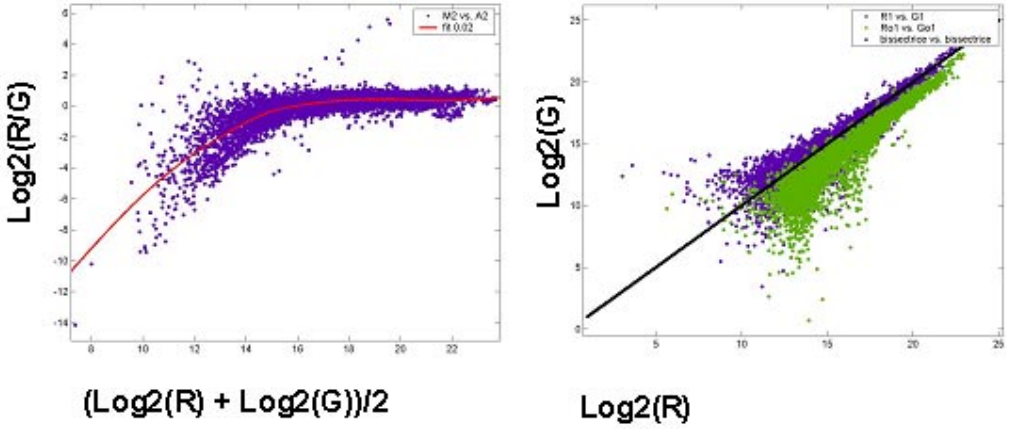


Figure 3.8. (A) Representation of the log intensity ratio M = log2 R/G versus the mean log intensity A = (log2(R)+log2(G))/2. At low average intensities on average the ratio becomes negative indicating that the green dye is consistently more intense than the red dye. Either the sensitivity of the red signal is lower than the one of the green signal or the basal noise level on the green signal is more pronounced. To compensate for nonlinear dye effects a Lowess fit with f value of 0.2 was used (solid line). (B) x axis log2(R); y axis: log2(G) Represents a plot of the log intensity of the R versus G prior to normalization by Lowess (green dots) and after Lowess normalization (blue dots). (New values of the log intensities of red and green were calculated based on the Lowess fitted values M and A.)

In summary, after these transformations, we can use the preprocessed values (corrected for array, spot, dye, and condition effects) to identify which genes show a differential level of expression between the two conditions by using a test statistic (T-test, paired T-test, Bayesian T-test). The drawback of most of these classical test statistics is the need for a high number of replicates. Since microarray experiments are expensive and labor intensive, the limited number of replicates usually available decreases the reliability of using classical T-tests.

III.3.5. Analysis of variance

An alternative approach that avoids the use of ratios and the need for a high number of replicates is based on analysis of variance (ANOVA). Summarizing, ANOVA models the measurement of the expression level of each gene as a linear combination of the major sources of variation (cfr. above). Several major effects representing the condition, dye and array effects, and combinations (2-, 3-, and 4-level combinations) of these main effects are taken into account in the models. Not all of the combined effects, however, have a physical meaning and only those considered to be important are retained in the models. Reliable use of an ANOVA model therefore requires a good insight into the process to be modeled. The model that best corresponds to physical reality is preferred. The most important combined factor in all models is the *GC* effect, the factor of interest. The *GC* effect reflects the expression of a gene depending on the condition (i.e., condition specific expression). The following equation represents the ANOVA model of microarray data taking into account the gene effect (G), condition effect (C), the dye effect (D), array effect (A), replicate effects (R), spot effects (AG):

$$I_{ijklm} = \mu + G_i + C_j + A_k + D_l + R_{m(i)} + (AG)_{ki} + (GC)_{ij} + \varepsilon_{ijklm}$$

The effect of interest (GC), R is 'nested' within G (indicated by brackets in the subscripts).

One of the basic requirements of ANOVA is that all dependencies need to be linear. This should be reflected by a normal distribution of the residuals of the fit with zero mean and equal variance. These requirements imply that mathematical transformations (log transformation described above) are mandatory to compensate for multiplicative effects. The prerequisite of normally distributed residuals however, is not too stringent; a proper ANOVA analysis can be done when residuals are independent and identically distributed (*i.i.d.*), but not necessarily normal, with zero mean and constant variance.

One of the major advantages of the ANOVA approach is that it allows gaining more information from the data than by looking at each gene separately (e.g., the array effect is similar to all genes on an array). Since all measurements are combined to allow statistical inference, the need for a high number of replicates is less pronounced. Figure 3.9 shows the results of an ANOVA model on Lowess normalized data that takes into account the 4 main factors (see above) and the factor of interest. It compensates for array, dye and condition effects and spot effects. To model the spots, a relationship between spots on the same array and a relationship between all left and right spots is assumed.

Source	SS	df	MS
G-effects	133176.7	3784	35.2
C-effects	0	1	0
A-effects	22432.4	1	22432.4
D-effects	0	1	0
AG-effects	9052.6	3784	2.4
RG-effects	2164.7	3785	0.6
GC-effects	1239.8	3784	0.3
Error	5656	15139	0.4
Corrected Total	173722.2	30279	5.7

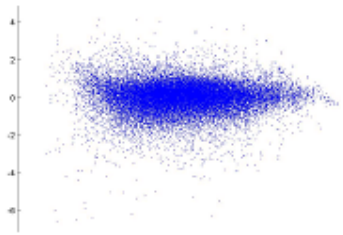


Figure 3.9. Result showing the ANOVA table and corresponding residual plot of the ANOVA model on the log transformed Lowess normalized data.

An important problem with current ANOVA models for microarrays that nonlinearities can be observed in the residual plots, which means that the basic assumptions of the model are not entirely fulfilled. Finding a proper data transformation that would render the residuals linear with respect to the estimated intensity values is the current focus of our research.

Obviously, microarray experiments are not always simple black & white experiments, like the one explained above. Very often more complex designs are used to investigate biological processes of interest. This is certainly true when more than two conditions need to be compared (e.g., a time course experiment, a comparison between different mutant strains, and so on). The methodologies just described can be extended without much trouble to handle these more complex experiments.

It is clear that microarray technology is very powerful but data generated by these techniques must be handled with caution. At first, there is a need for a consistent recording of the complete production procedure as such that mistakes can easily be traced. Secondly, data need to be cleaned prior to further analysis. Once preprocessed, the data can be used for further data exploration and data mining.

III.4. Microarrays and disease management

After discussing in the last three sections how microarray information can be produced, managed, and transformed into reliable measurements, we now open the door to more complex analyses. This section demonstrates how microarrays contribute to the disease management goals presented in Chapter 2 and how data mining can be applied to such data. We show that microarray measurements can be interpreted as a molecular fingerprint of a disease and we present a data-mining framework where this information can be used to improve disease management.

III.4.1. Molecular fingerprinting of malignancies

Although the exact mechanisms of carcinogenesis are usually unknown, it is generally assumed that most cancers originate from genetic disorders (mutations in proto-oncogenes, tumor suppressor genes) induced by carcinogens, viral infections, and so on). If these mutations induce changes in the expression levels of other genes (genes without mutations, but of which the expression levels are directly or indirectly controlled by the mutated genes) the measured expression levels will be disturbed as compared to those of cells derived from healthy tissues.
Moreover, if the class distinction between different tumor types finds its origin in different mutations and genomic aberrations (Sager, 1997), the induced perturbations in the observed expression profiles are tumor specific and serve as a molecular fingerprint of a tumor belonging to a certain class.

For cancer, expression profiles can be measured for tumor cells coming from different classes, for example:

- Tumors with a different histopathological diagnosis.
- Tumors in a different stage of development or different staging.
- Tumors with a different effect of therapy.
- Tumors with a different prognosis (e.g., disease-free survival versus relapse for tumors without clinical apparent metastasis or with and without subclinical metastases; indolent versus more aggressive disease).
- Benign versus malignant tissue.
- Primary tumor versus metastasis.
- Tumor taken before and after chemotherapy.

The fingerprint of a certain tumor type as determined by microarray analysis, consists of a vector with thousands of components (1 component = expression level for a gene). Fingerprints of distinct tumors and cell lines (samples) can be compared by arranging the vector corresponding to each fingerprint in an expression matrix (the rows contain the expression levels of a specific gene and the columns contain the expression levels of a specific sample or tumor). The high dimensionality of these datasets necessitates the use of specific methods to extract clinically useful information.

In this context a data-mining framework (De Smet et al., 2001) was developed that uses the expression matrix and specific algorithms to facilitate diagnosis, staging, prognosis estimation, therapy planning, and follow-up and to obtain a more fundamental insight in the molecular biology of carcinogenesis.

III.4.2. Development of a data-mining framework

In the next paragraphs we will give an overview of this framework. As an example we have analyzed, using our framework, a data set containing 72 patients with acute lymphoblastic leukemia (ALL, in this text we say that these patients belong to class 1) or acute myeloid leukemia (AML, in this text we say that these patients belong to class 2). Peripheral blood or bone marrow samples of these patients were analyzed with an Affymetrix oligonucleotide array (with about 7000 probes) (Golub et al., 1999). These measurements (7129 x 72 matrix) are publicly available (http://waldo.wi.mit.edu/MPR/data_set_ALL_AML.html). The 72 patients were divided in two groups:

1. The training set: containing the first 38 patients (27 with ALL and 11 with AML).
2. The test set: containing the remaining 34 patients (20 with ALL and 14 with AML).

We will refer to the entire data set (no distinction between training and test set) as the leukemia data set.

III.4.2.1. Feature selection

Not all 7129 genes in this example are correlated with the class distinction between the samples and one is interested in discovering those genes that are truly relevant for the discrimination between the samples (i.e., feature selection). Feature selection is a way of reduction of the dimensionality of the problem (e.g., to reduce the number of dimension from 7129 to 5 in the leukemia data set). Feature selection can be done in two ways:

Selection of individual genes

This is the most straightforward way to do feature selection. Here we want to identify single genes whose expression is correlated with a class distinction known in advance (notice the difference with class discovery where the class distinction is not known in advance).
For this simple method of feature selection, a measure of correlation P(g) introduced by Golub et al. (1999) was used:

$$P(g) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) - \sigma_2(g)},$$

where $\mu 1$ and $\mu 2$ are, respectively, the mean values of the expression levels of gene g belonging to samples from class 1 and 2 - $\sigma 1$ and $\sigma 2$ are the associated standard deviations.
As an example, we calculated P(g) corresponding to the class distinction ALL-AML for all the 7129 genes in the leukemia data set. We selected the 5 genes with the highest absolute value of P(g). They are displayed in Table 3.1.

Gene description	P(g)
CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	-1.5956
CTPS CTP synthetase	1.5494
Leukotriene C4 synthase (LTC4S) gene	-1.4959
DF D component of complement (adipsin)	-1.3935
C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds	1.3719

Table 3.1. Five genes with highest discrimination between ALL and AML samples.

Retaining only the expression levels for these genes is equivalent with a dimensionality reduction from 7129 to 5 and leaves us with 5 genes that might be at the origin of the molecular biological difference between ALL-AML.

Selection of a combination of genes

A second, more complicated method to reduce the dimensionality is the selection of different combinations (linear or non-linear) of gene expression levels. Each of these combinations can be regarded as one feature. Note that in the previous method each feature corresponds to exactly one gene expression level. However, in general, the distinction between classes is not fully determined by the activity of a single gene, but rather by the interaction of several genes. It is therefore better to work with a combination of genes.

One of the methods we currently use in this context is Principal Component Analysis (PCA) (Bishop, 1995). Shortly summarized, we can say that PCA finds linear combinations of gene expression levels in such a way that these combinations have maximal spread (or standard deviation) for a certain collection of microarray experiments. In fact, PCA searches for the combinations that are most informative. These (linear) combinations are called the Principal Components for this collection of experiments.

In the leukemia data set (m=7129), we have calculated the two (p=2) first Principal Components of the training set (n=38) and determined the associated features (two for each sample) for the samples in the training and test set. The results are shown in Figure 3.10.

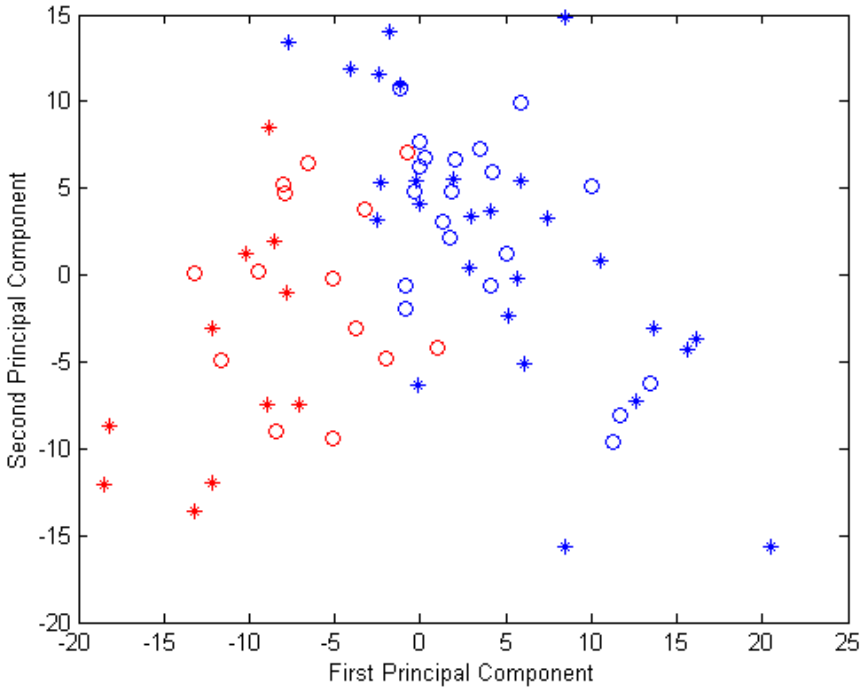


Figure 3.10. Principal Component Analysis for the leukemia data set. First, we calculated the first two Principal Components of the training set. Secondly, we projected every sample from the training AND test set onto these two Principal Components, resulting in two values for each sample (in this figure plotted in the X- and Y-axis). Training set: *=ALL, *=AML; Test set: O=ALL, O=AML.

Note that for this feature selection method the class distinction does not have to be known in advance, making this procedure suitable to be used in combination with class discovery.

III.4.2.2. Class prediction

In a clinical environment it is important to be able to do predictions (with regard to histopathology, staging, prognosis and therapy choice), using microarray experiments, for individual patients. This is the problem of class prediction. Here a prediction must be made for samples or patients for which class membership (like the response to therapy, stage, …) is not known in advance. Based on a set of features and a training set (experiments done in the past - for which class membership is already known), a model (like Neural Networks, Support Vector Machines, Linear Discriminant Analysis, …) has to be trained. This model can then be used to classify new patients (prediction of class membership). This will constitute a true revolution in the management of cancer (compare with the classical, rather empirical approach) because in the future, microarrays will potentially be available to many physicians and will therefore be a common tool leading clinical management of cancer (Friend, 1999).

For the leukemia data set, we used Linear Discriminant Analysis (Fisher) (Bishop, 1995) to construct a linear model after Principal Component Analysis (as described in the previous paragraph and Figure 3.10). We used the training set to determine the parameters of this model (training – determination of a line in the Figure above that optimally separates the patients of the training set in ALL (*) and AML (*) patients). Then, we used this model to classify the patients of the test set, which resulted in 3 misclassifications. This is shown in Figure 3.11.

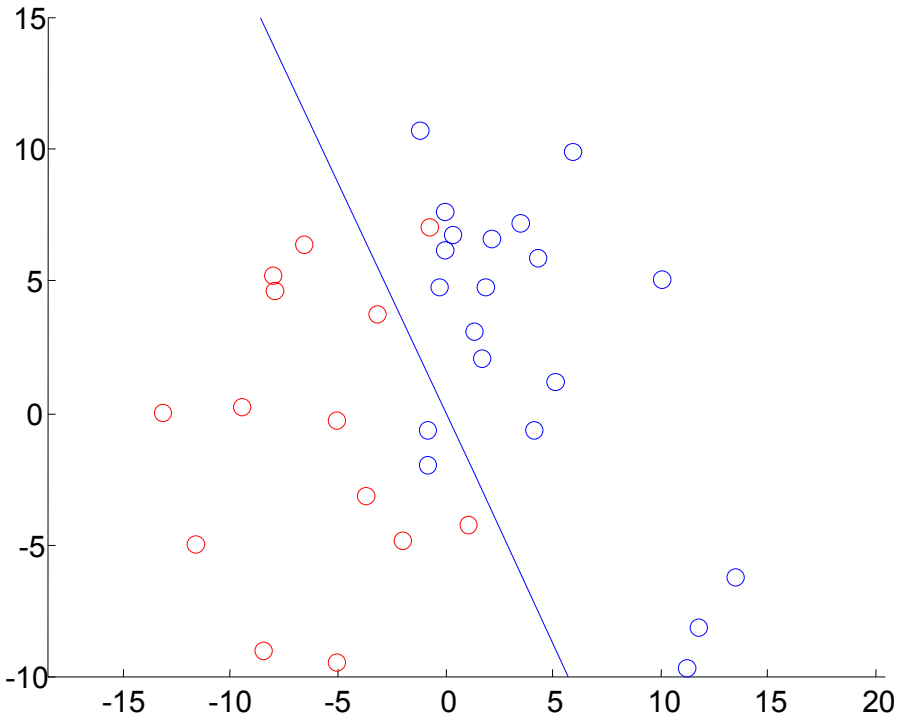


Figure 3.11. Model for the classification of patients with acute leukemia after Linear Discriminant Analysis using the first two (p=2) Principal Components of the training set. The parameters of the linear model (blue line) were calculated using the patients of the training set (* and * in the Figure 3.10). The patients (only the patients of the test set are shown in this figure) above the line are classified as ALL and below as AML. Note that this results in 3 misclassifications. Test set: O=ALL, O=AML.

Through the use of more Principal Components (e.g., p=5) or the use of non-linear models (Support Vector Machines), we were able to further reduce the number of misclassifications.
The procedure described above, could be applied to develop models (using measurements with microarrays) to predict other aspects of the clinical course of a malignant process.

- Models could be develop to predict the presence or absence of subclinical metastases, which would allow us to select the patients that would benefit from adjuvant therapy and also select the patients for which adjuvant (chemo-)therapy would mean unnecessary toxicity.
- Models could be developed to make an estimate of the prognosis. Many tumors, that appear clinically similar, can however have very different clinical outcomes (indolent versus aggressive disease; for example, for malignancies of the prostate). These differences, however, can show themselves on a molecular level making them visible for analysis with microarrays.
- Models could be developed to predict the therapy response using cells of the untreated tumor. These models could be used to select the most efficient therapy.

III.4.2.3. Cluster analysis of patients or microarray experiments

Using the extensive arsenal of clinical and morphological parameters, malignant processes can be divided in different diagnostic categories or entities with similar clinical behavior. In most cases, these categories guide the clinical management. It is however possible that rearranging the diagnostic categories might result in groups of patients with less clinical variability, which would allow us to refine the clinical management.

With cluster analysis it is possible to automatically find different classes or clusters in a group of microarray experiments without knowing the properties of these classes in advance. A cluster, in general, will group microarray experiments (or the associated patients) with a certain degree of similarity. The distinct classes or clusters generated by the clustering procedure will probably - at least partially – match with the existing diagnostic categories used for the current classification tumors. However, it is not excluded that novel, yet unknown diagnostic entities might originate from these analyses which would improve clinical management of cancer (Alizadeh et al., 2000) (Perou et al., 2000).

These techniques can be illustrated using the patients of the entire (here there is NO subdivision between training and test set) leukemia data set. Now imagine that the difference between ALL and AML is not known. In this case we have simply a data set with 72 patients with acute leukemia. We have clustered these 72 experiments or patients using two well-known clustering procedures: K-means (Tou and Gonzalez, 1979) as presented in Figure 3.12 and hierarchical clustering (Eisen et al., 1998) as presented in Figure 3.13.

Figure 3.12 shows that K-means clustering succeeded in finding two clusters. In the first cluster, all the patients – except one – have ALL. The second cluster contains patients having AML. This means that the procedure was in fact able to redefine the concepts ALL and AML.

We also performed hierarchical clustering on the patients of the entire leukemia data set since this is the method that is commonly used in literature. No Principal Component Analysis is necessary before starting the procedure. This technique iteratively groups the two most similar (according to a certain distance measure like the Euclidean distance or the Pearson correlation coefficient) elements in the data, creating a tree structure. Dependent on the level where the tree is cut, clusters are defined. The result, using the patients of the leukemia data set, is presented in Figure 3.13. Note that the largest part of the patients with AML is grouped in the right side of the tree structure (see the branch where the arrow is pointing).

In this example, nothing new is learned (because ALL and AML were already known). Both cluster methods identify the distinction between ALL and AML as the most pronounced grouping in the data set but the result clearly shows the potential of this technique for novel class discovery. The finer the class distinction is the more the treatment of the disease can be fine tuned and individualized. Some class differences between disease types might only be discernable at the molecular level. Therefore class discovery can greatly enhance optimal disease management.

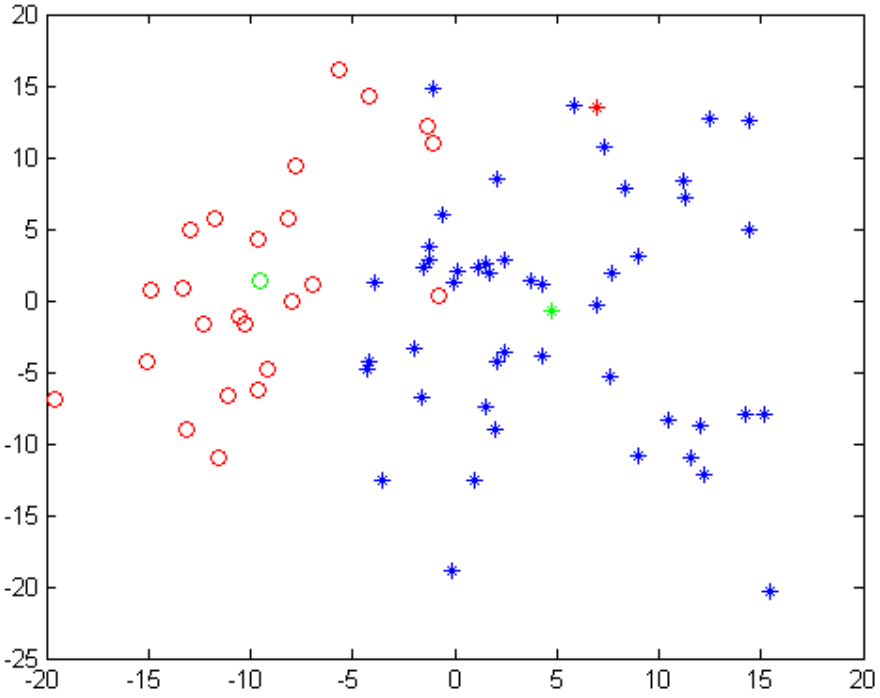


Figure 3.12. K-means clustering with two clusters for the patients with acute leukemia after Principal Component Analysis according to the first five Principal Components of the entire leukemia data set. Only the first two Principal Components are shown in this figure. Note the almost perfect correlation between the clusters and the clinical classification (ALL-AML). Cluster 1: *=ALL, *=AML, *=cluster mean; Cluster 2: O=ALL, O=AML, O=cluster mean.

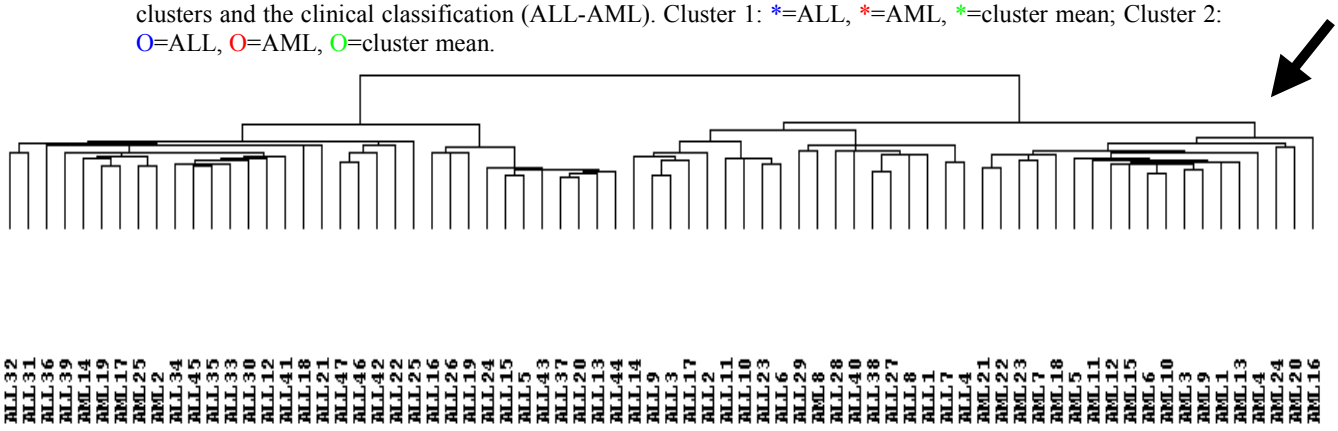


Figure 3.13. Hierarchical clustering for the patients of the leukemia data set. The terminal branches represent the individual patients (ALL/AML + number). Note that most patients with AML (19 out of 25) are grouped in the branch where the arrow is pointing.

III.5. Conclusion

In this chapter, we introduced microarray technology, which is one of the recent technologies that are having a major influence on research in molecular biology. First, we introduced the basics of microarray technology that relies on measuring gene expression by hybridization of gene mRNA transcripts on miniaturized DNA probes. We then described how a Laboratory Information Management System keeps track of all the steps necessary for the deployment of microarray technology. Next, we discussed the preprocessing of the raw data from microarray experiments. As these experiments have a high level of noise, it is essential to clean up the data to obtain reliable measurements. To this end, we present two classes of techniques: normalization and analysis of variance (ANOVA).

Second, as an illustration of the principles of microarray data analysis and to make clear how microarray technology fits within the clinical themes from the previous chapter, we show how to use microarray data in building models for disease management in oncology. At the hand of microarray data from two types of leukemia, we illustrate the different tasks in this area: (1) the selection of the features most relevant to some clinical outcome, (2) the prediction of the clinical outcome using statistical models similar to those presented in the previous chapter, and (3) the discovery of classes of malignancies at the molecular level (which could be different from the current medical classification and could provide more insight into the behavior of the malignancy) by clustering algorithms.

IV. From expression to regulation

As said previously, microarray experiments generate a considerable amount of data, which analyzed properly help us gain a huge amount of biologically or medically relevant information about the global cellular behavior (e.g., normal versus tumor cells, dividing cells in the cell cycle). In this chapter we focus on the exploratory analysis of the microarray data to gain deeper insights in the fundamental biological processes being at the origin of complex disease.

One of the first steps in data analysis of high-throughput expression measurements (after pre-processing) is the clustering of genes to find groups of genes that have a similar behavior or expression profile. We describe a specialized algorithm, called adaptive quality-based clustering, that was developed for the specific task of finding coherent groups of genes. Another way to explore the microarray data is by looking at the groups of genes that have a common functional annotation and by correlating the expression patterns with the functional classes. We have developed a set of web-based tools to browse the gene ontology and visualize the corresponding gene expression data. The next step involves the inclusion of sequence information in the analysis. Once interesting groups of genes are identified we look at the DNA sequence of these genes. In this case we are particularly interested in the region upstream of the gene where the transcriptional regulators bind to the DNA. We present the analysis of the activation of a known transcription binding *PLAG1*, which can result in the uncontrolled activation of downstream target genes and identify likely targets of this transcription factor. We also present Gibbs sampling, which makes it possible to identify, through a stochastic search method, possible motifs in upstream regions when the motif we are looking for as never been identified before. Finally, we end this chapter with an example of the integration of the described tools.

IV.1. Clustering gene expression profiles

Currently analysis of microarray data starts mostly with clustering. A variety of techniques have been implemented and described: Self-Organizing Maps (Tamayo et al., 1999), hierarchical clustering (Carr 1997; Eisen et al., 1998), Self-Organizing Tree Algorithm (Herrero et al., 2001), K-means (Tavazoie et al., 1999; Tou and Gonzalez, 1979), simulated annealing (Lukashin and Fuchs, 2001), Principal Component Analysis (Quackenbush, 2001), Multidimensional Scaling (Bittner 2000), Cluster Affinity Search Technique (Ben Dor et al., 1999). One of the objectives of these methods is to detect groups of genes that exhibit a highly similar expression profile. Since gene expression profiles are encoded in real vectors, these algorithms intend to group gene expression vectors that are sufficiently close to each other (according to a certain distance or similarity measure).

Most clustering algorithms originate from non-biologically related research fields. Therefore, although useful, the original implementations suffer from some drawbacks as has been highlighted by Sherlock, 2000. These deficiencies can be summarized as follows. Firstly, algorithms such as K-means and Self-Organizing Maps require the predefinition of the number of clusters (parameter of the algorithm). When clustering gene expression profiles, the number of clusters present in the data is usually unknown. Changing this parameter usually affects the final clustering result considerably. Clustering, using for example K-means, therefore involves extensive parameter fine-tuning to detect the optimal clustering and the choice of the final parameter setting remains somehow arbitrary (e.g., based on visual inspection of the clusters). When using hierarchical clustering, the number of clusters is determined by cutting the tree structure at a certain level. The resulting cluster structure is therefore highly influenced by the choice of this level, which in turn is rather arbitrary. Secondly, the idea of forcing each gene of the data set into a cluster is a significant drawback of these implementations. If genes are, despite a rather low correlation with other cluster members, forced to end up in one of the clusters, the

average profile of this cluster is corrupted and the composition of the cluster becomes less suitable for further analyses (such as motif finding, see further).

Much effort is currently being done to adapt clustering algorithms towards the specific needs of biological problems. In this context the ideas of quality-based clustering (Heyer et al., 1999) and gene shaving (Hastie et al., 2000) were developed. Heyer et al. proposed an algorithm (called QT_Clust) that tries to identify clusters that have a certain quality (representing the minimal degree of co expression needed) and where every cluster contains a maximal number of points. Genes not exhibiting this minimal degree of co expression with any of the clusters are excluded from further analysis. A problem with the quality-based approach of Heyer et al., however, is that this quality is a user-defined parameter that is hard to estimate (it is hard to find a good trade-off or optimal value: setting the quality too strictly will exclude a considerable number of co expressed genes, setting it too loose will include too many genes that are not really co expressed). Moreover, it should be noted that the optimal value for this quality is, in general, different for each cluster and data set dependent.

IV.1.1. Adaptive quality-based clustering

We have developed an adaptive quality-based clustering method (De Smet et al., 2002) starting from the principles described by Heyer et al. (quality-based approach; locating clusters, with a certain quality, in a volume where the density of points is maximal). The algorithm is in essence a heuristic, two-step approach that defines the clusters sequentially (the number of clusters is not known in advance, so it is not a parameter of the algorithm). The first step locates a cluster and the second step derives the quality of this cluster from the data. Table 4.1 gives a summary of the differences between our method, hierarchical clustering, SOM and K-means.

	Adap_Cluster	**Hierarchical clustering** (Eisen, 1998)	**SOM** (Tamayo, 1999)	**Standard K-means** (Tou, 1979)
Format of result	Set of clusters	Tree structure (difficult to interpret)	Set of predefined number of clusters	Set of predefined number of clusters
PRINCIPAL USER-DEFINED PARAMETER	Significance level S	-	Number of clusters / Node topology	Number of clusters
Additional requirements f rom the user	Limited (fine-tuning of S is rarely necessary)	Definition of (an arbitrary) level where the tree structure has to be cut	Extensive parameter fine-tuning (comparison of several runs with different parameter settings)	Extensive parameter fine-tuning (comparison of several runs with different parameter settings)
Statistical definition of clusters	Yes	No	No	No
INCLUSION OF ALL GENES IN CLUSTERS	No	Yes	Yes	Yes
Missing values management	Yes	Yes	Not discussed	Not standard
Computational complexity	Linear in N	Quadratic in N	Linear in N	Linear in N

Table 4.1. Comparison between adaptive quality-based clustering, hierarchical clustering, Self-Organizing Maps, and K-means.

Clusters formed by our algorithm might be good 'seeds' for further analysis of expression data (Thijs et al., 2002) since they only contain a limited number of false positives. When the presence of false positives in a cluster is undesirable, a more stringent value for the significance level S

might be applied (e.g., 99%; for noise-sensitive analyses such as motif finding) which will result in smaller clusters exhibiting a more tightly related expression profile.

The algorithm is a heuristic iterative two-step approach. There are two user defined parameters: (1) the minimal number of genes in a cluster and (2) the significance level S.
During each iteration, this algorithm first finds a cluster center using a preliminary estimate of the radius or quality of the cluster. Given a collection G of gene expression profiles, the objective is to find a cluster center in an area of the data set where the 'density' (or number) of expression profiles, within a sphere with an initial radius or quality equal is locally maximal. After initialization of the cluster center (with the mean profile of all the expression profiles in the data set G), all the expression profiles within a sphere with radius RAD are selected. Iteratively, the mean profile of these expression profiles is calculated and subsequently the cluster center is moved to this mean profile. This approach moves the cluster in the direction where the 'density' of profiles is higher. This is conceptually visualized in 4.1.

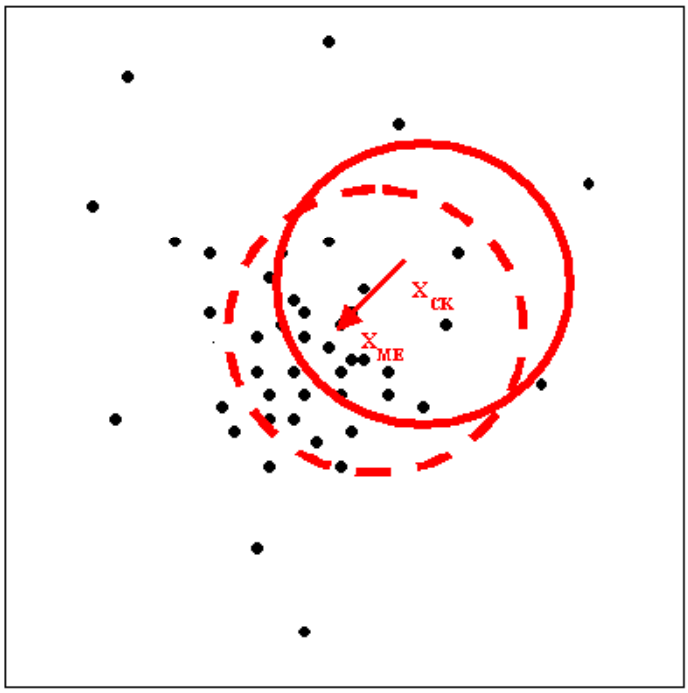


Figure 4.1. Conceptual representation of the relocation of the cluster center.

When the cluster center has been located, it remains fixed. The algorithm now determines a new estimate for the radius of the cluster. To estimate the radius of the cluster, first the distance of each gene in the data set to the cluster center is calculated. The distribution of these distances in the original data consists of two parts (see also Figure 4.2):

1. Genes belonging to the background of the current cluster center, which do not belong to any, cluster (noise; are not significantly co expressed with other genes) or belong to another cluster. The distribution of the distance of these genes does not differ significantly from the distribution of a randomized data set (distribution shown in red in Figure 542).
2. Genes belonging to the cluster have a distribution that is not present in the distribution of the randomized data set (left sided blue tail in Figure 4.2) and are significantly co-expressed.

Finding the parameters of the proposed model is accomplished by fitting the model to the distance distribution with the EM-algorithm.

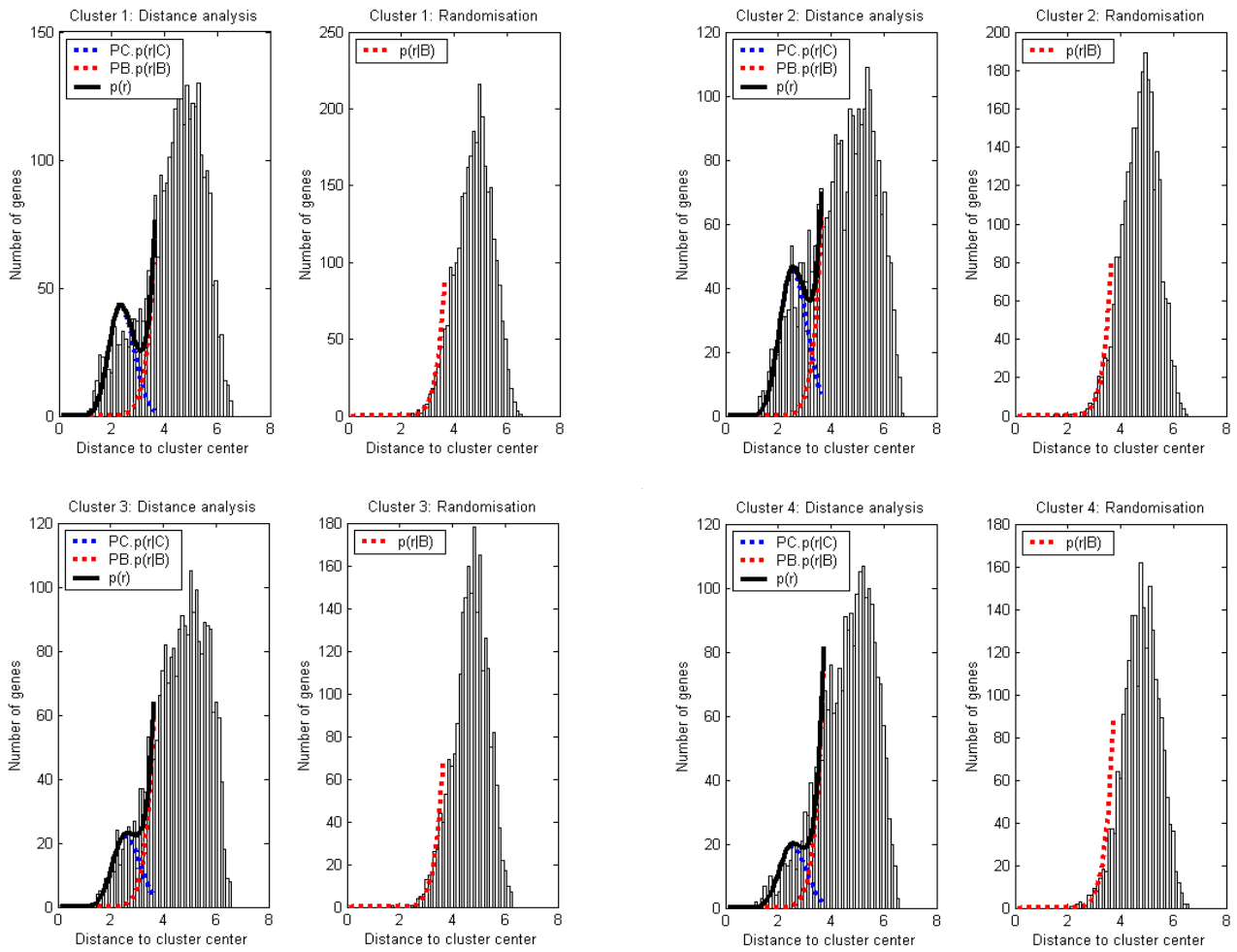


Figure 4.2. Distribution of the distance from the cluster center for all points in the data set for the first 4 clusters in the Cho data set (see below).

We have built a web interface to access the algorithm online. Figure 4.3 is a screenshot of part of the results page. First an overview of the parameters is given. For each cluster a plot of the expression profiles is given together with a list of all the genes present in the cluster. In this example the genes are identified by their accession umber and gene name.

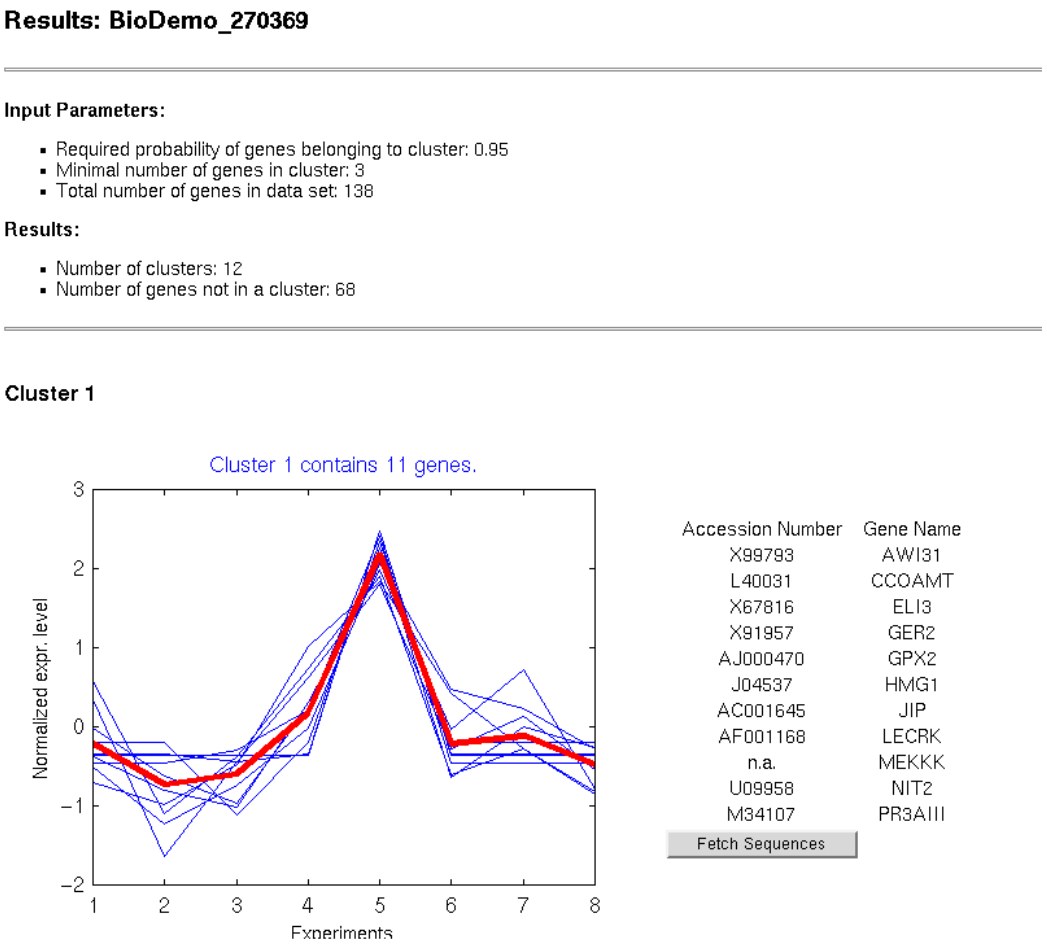


Figure 4.3. Screenshot from the results page of the clustering web interface. The normalized expression profiles are shown together with the identifiers of the genes present in the cluster.

IV.1.2. Clustering gene expression in the mitotic cell cycle of yeast

As an example, the algorithm was tested on the expression profiling experiment of Cho et al., 1998 studying the yeast cell cycle in a synchronized culture on an Affymetrix chip. The cell cycle of yeast is a fundamental biological system as it reveals the core processes involved in cell replication and growth in general. This knowledge is essential to the understanding of the aberrant processes involved in tumorigenesis and carcinogenesis. This data set can be considered as a benchmark and contains expression profiles for 6220 genes over 17 time points taken at 10-min intervals, covering nearly two full cell cycles. The majority of the genes included in the data set have been functionally classified (Mewes et al., 2002), which makes this data set an ideal candidate to correlate the results of new clustering algorithms with the biological reality.

Our pre-processing included the following steps: (1) data corresponding to the 90 and 100-min measurements were removed (Tavazoie et al., 1999), (2) the 3000 most variable genes were selected and (3) the gene expression profiles were normalized. The main results of the cluster

analysis with our algorithm (minimal number of genes set to 10 and the significance level to 0.95) are summarized in Table 4.2

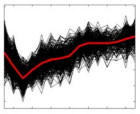
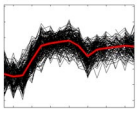
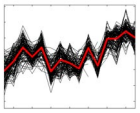
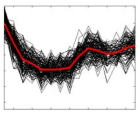
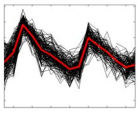
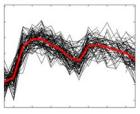
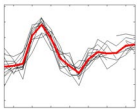
Cluster number	Graphical representation of cluster	Number of ORFs	MIPS functional category (top-level)	ORFs within functional category	P-value (-log$_{10}$)
1		426	energy transport facilitation	47 40	10 5
3		196	cell growth, cell division and DNA synthesis	48	5
4		149	protein synthesis cellular organization	71 107	50 19
5		159	cell rescue, defense, cell death and ageing	20	4
6		171	cell growth, cell division and DNA synthesis	76	24
9		78	cell growth, cell division and DNA synthesis	23	4
37		11	metabolism	9	6

Table 4.2. Main results of the clustering algorithm.

For the biological validation of this result we looked for enrichment of top-level functional categories in individual clusters. We mapped the genes in each cluster to the top-level functional categories in the Munich Information center for Protein Sequences (MIPS) Comprehensive Yeast Genome Database (Mewes et al., 2002). For each cluster we calculated P-values for observing the frequencies of genes in particular top-level functional categories using the cumulative hypergeometric probability distribution (Tavazoie et al., 1999). Note that we were able to determine the role of every cluster presented in Table 4.2. within the yeast cell cycle context and to correlate this role with the behavior of the average profiles of the clusters. We have also found several protein complexes where nearly all members belonged to the same cluster.

IV.2. Using functional annotations of genes: Gene Ontology

In the previous section we discussed the use of clustering algorithms to find groups of genes that have a similar expression profile. Another possibility to explore the microarray data is to look at interesting functional classes of genes. Therefore we use the largest public functional information database developed by the Gene Ontology consortium (Ashburner et al., 2000). In this database functional information is represented as a tree of nested functional categories. At the top level the database is separated into three "root" categories (called ontologism) of biological process, molecular function and cellular component. The sub-terms of these terms and the genes associated with them can easily be explored at several GO websites (e.g., AmiGo at http://toy.lbl.gov:8888/cgi-bin/go.cgi, see Figure 4.4).

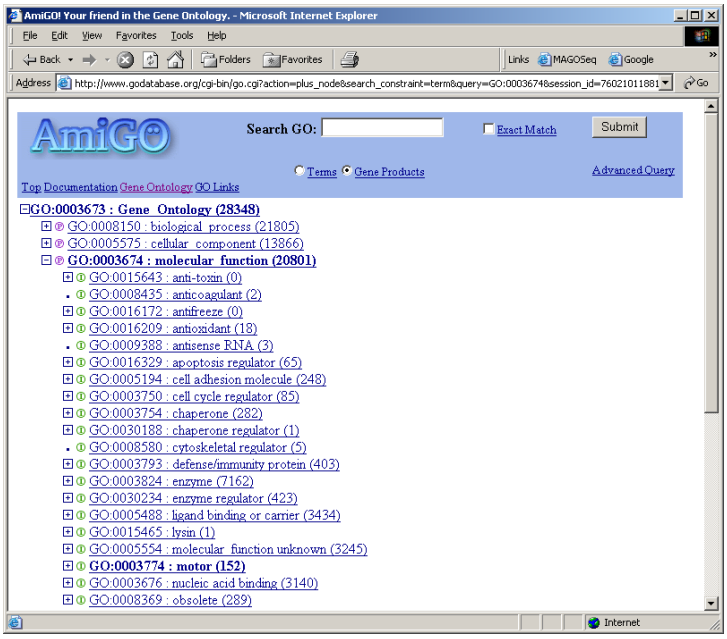


Figure 4.4. A subset of the *molecular function* gene ontology tree as can be viewed with AmiGO or any other Gene Ontology browser that is available on http://www.geneontology.org.

Each gene can be associated with one or many functional categories (GO terms) of this tree. Associations of a gene with a particular term A makes this gene automatically associated with a (parent) term B that contains term A, and so on, all the way up to the root.
Some GO terms do not contain all the relevant genes that can be found in the literature. To make up for this we made our own lists of genes involved in several functional processes, based on the literature. The advantage of using the controlled functional vocabulary defined by GO is huge, and the annotation quality improves every month. Nonetheless, biologists interested in expression patterns of genes within their field of expertise will always be ahead of database or even literature annotation. Therefore we have implemented an interface permitting the creation of user-defined gene sets, by searching with keywords or gene identifiers, and also with lists of either of them.
Expression data for genes associated with any GO term or for genes of a gene set can easily be accessed, clustered, visualized and downloaded (in graphical or text format) from a web application we have developed, namely MAGOSeq. This is an acronym for MicroArrays, Gene Ontology and DNA Sequence analysis and implements the following analysis process for microarray data:

- Preprocess the raw data using a smoothing spline fit for each slide
- Filter the data according to the consistency of replicate measurements

- Store the preprocessed data in a relational database together with the identifiers and the functional annotation of the genes:
  - GenBank accession number (http://www3.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide)
  - Unigene identifier (http://www.ncbi.nlm.nih.gov/UniGene/)
  - Mouse Genome Informatics (MGI) identifier (http://www.informatics.jax.org/)
  - Ensembl ID (http://www.ensembl.org)
  - Locuslink ID (http://www.ncbi.nlm.nih.gov/LocusLink/)
  - Gene ontology associations (http://www.geneontology.org)
- Explore and analyze the expression profiles of different groups of genes
  - Construct heat maps (Figure 4.6) and profile charts (Figure 4.7)
  - Export text files for use with other analysis tools
- Export a description file for a gene group (in eXtended Markup Language or XML) that can be imported in the linked sequence analysis package (Figure 4.8)
- Retrieve DNA sequences for the upstream regions of the genes in a gene group and annotate these sequences with promoter regions (Scherf et al., 2000), CpG islands (Pedersen et al., 1999) and known transcription factor binding sites (Wingender et al., 1996)
- Select short fragments of these sequences to be analyzed with our Gibbs motif sampler.

The MAGOSeq system (Figure 4.5) integrates different bioinformatics techniques for the analysis of microarray data in one user-friendly application. More about such integration is discussed below. MAGOSeq is not yet publicly available but has been used extensively within our department for the analysis of a huge microarray experiment (with 60 DNA chips) in collaboration with molecular biologists of the faculty of Medicine of the K.U.Leuven.

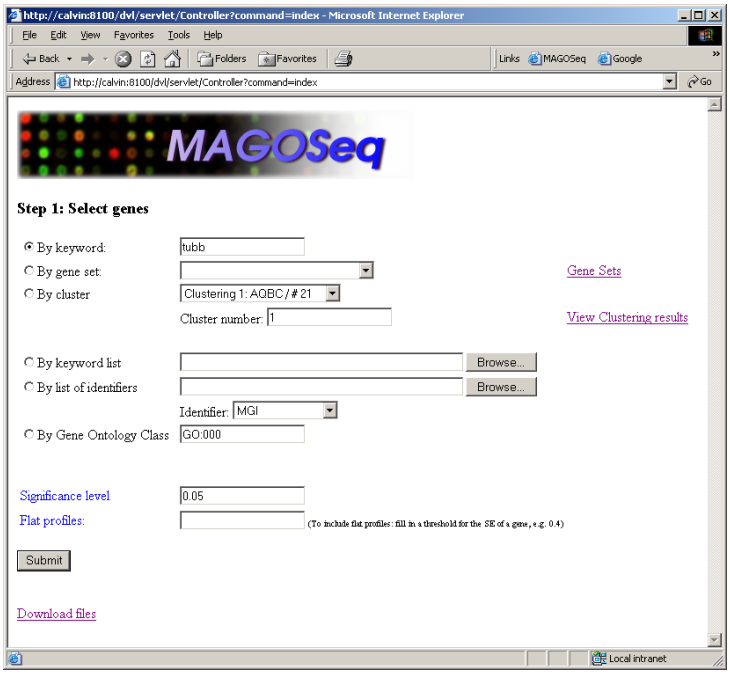


Figure 4.5. The MAGOSeq web application for the functional analysis of microarray data using Gene Ontology. Next to Gene Ontology classes, gene groups can also be constructed by selecting a gene cluster or by searching for genes with keywords or identifiers.

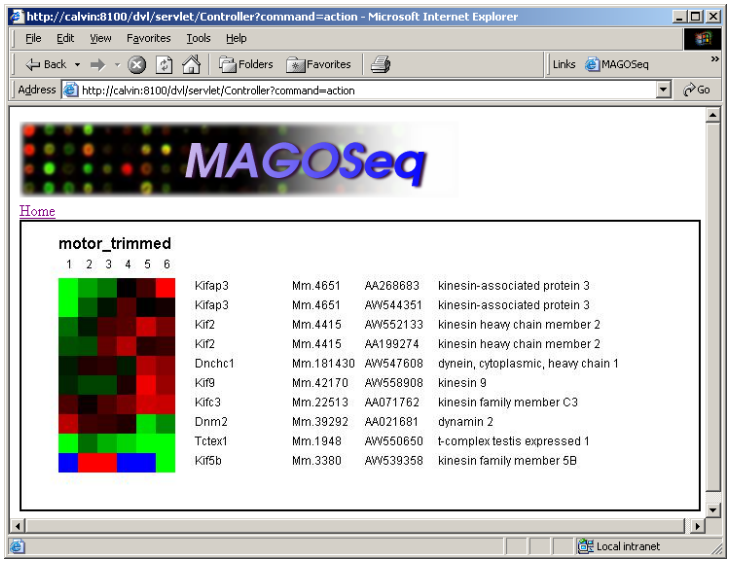


Figure 4.6. Representation of the expression profiles as a *heat map*. The genes in this group are all motor-related as defined by their Gene Ontology annotation. A green rectangle represents under expression as compared to a reference sample. A red color represents over expression, a black color a constant expression and a blue color is a missing value.

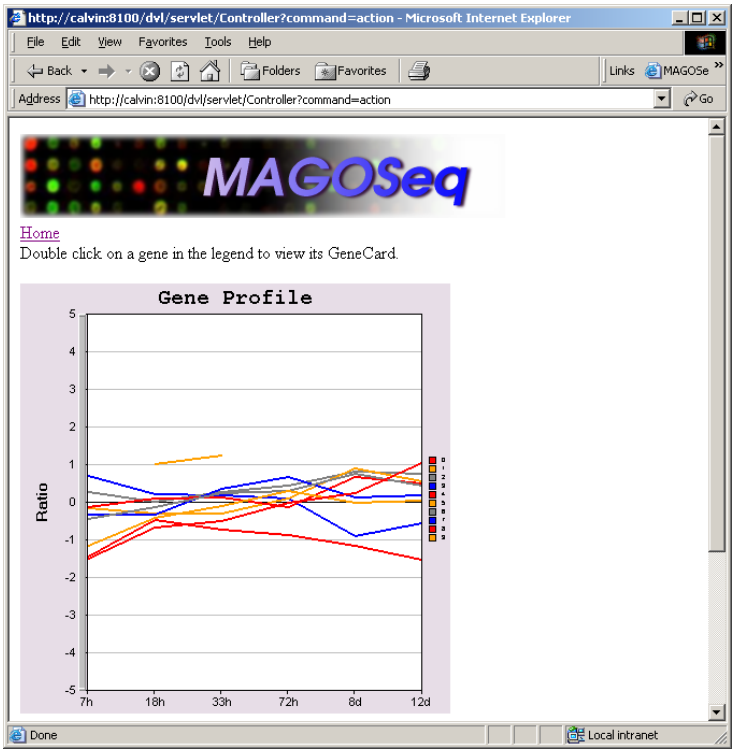


Figure 4.7. The same group of motor genes as in the Figure 4.6, this time represented in a profile chart. It represents the changes of expression of these genes over time (in logarithmic scale). As can be seen on the chart, some profiles have missing values. These were filtered out as being unreliable.

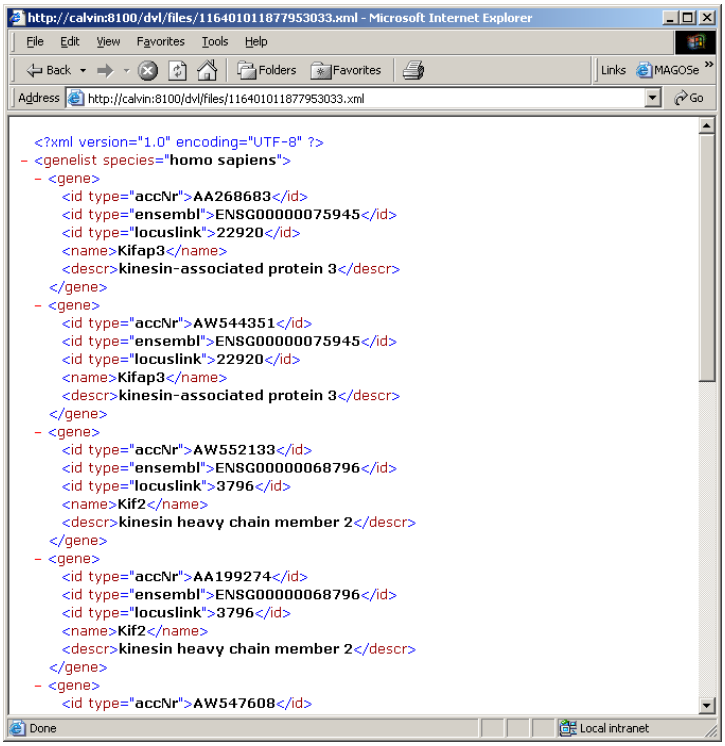


Figure 4.8. The name, description and several external database identifiers of the genes in a gene group can be represented in XML format. Such XML files are ideally suited for the exchange of information between different systems and thus they promote integration.

IV.3. Whole-genome scan for genes regulated by a known transcription factor

Once potentially interesting genes have been identified by analyzing the microarray measurements (by statistical tests of differential expression, clustering, and so on), we can then study these genes further at the sequence level. In this section, we search for potential targets of the transcription factor PLAG1 by searching for potential binding positions of this transcription factor in the upstream regions of genes that have been identified as potentially regulated by PLAG1 in a microarray experiment.

PLAG1 is a developmentally regulated C_2H_2 zinc finger protein that acts as a transcription factor by recognizing and binding DNA and/or RNA. The mammalian genome is known to contain a large number of zinc finger protein-encoding genes (Pieler and Bellefroid, 1994), and the number of such genes implicated in cancer is growing steadily. For instance, the activation of PLAG1 in pleomorphic adenomas of the salivary gland results in uncontrolled activation of downstream target genes, causing the formation of a tumor (Kas et al., 1997).

IV.3.1. Introduction

IV.3.1.1. Zinc finger proteins

The zinc finger is a major structural motif involved in protein-nucleic acid interactions. Zinc finger proteins represent the largest class of DNA binding proteins in eukaryotes and zinc finger protein-controlled gene expression is a fundamental aspect of development as well as a known cause of cancer. These proteins are called zinc fingers because Zinc (Zn) ions coordinate finger-like 3D structures through bonds with cysteine and histidine residues. Details on the structure of zinc fingers are shown in Figure 4.9.

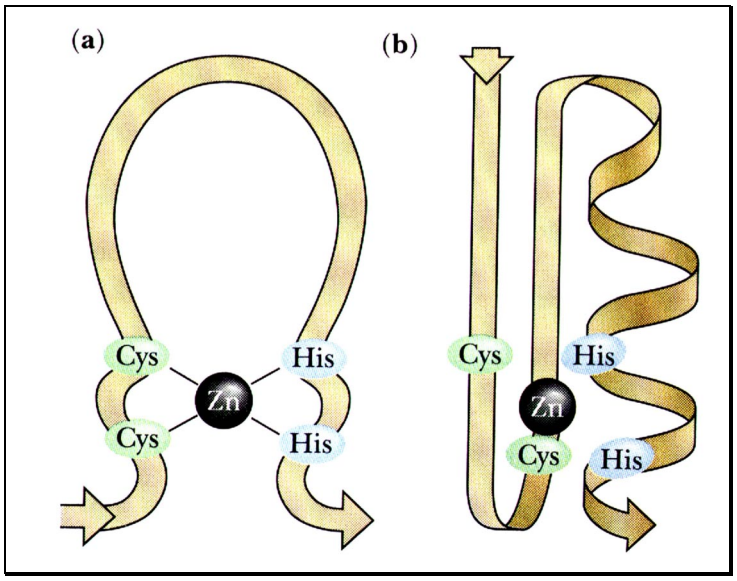


Figure 4.9. Zinc finger of the C_2H_2 type. (a) coordination of cysteine and histidine residues by zinc. (b) secondary structure of a C_2H_2 zinc finger. (Adapted from Evans R.M., Hollenberg S.M., 1988 (Cell 52:1, Figure 1)).

Each zinc finger functions as an independent module, binding to three base pairs in the DNA, which is shown in Figure 4.10. The C_2H_2-type zinc finger motif defines an extremely large family of DNA/RNA-binding proteins with probably more than 1000 structurally distinct members

present in the human genome. Apart from transcriptional modulation and control of RNA metabolism, chromatin packaging might also be an important activity through which zinc finger proteins exert their regulatory roles (Pieler and Bellefroid, 1994).

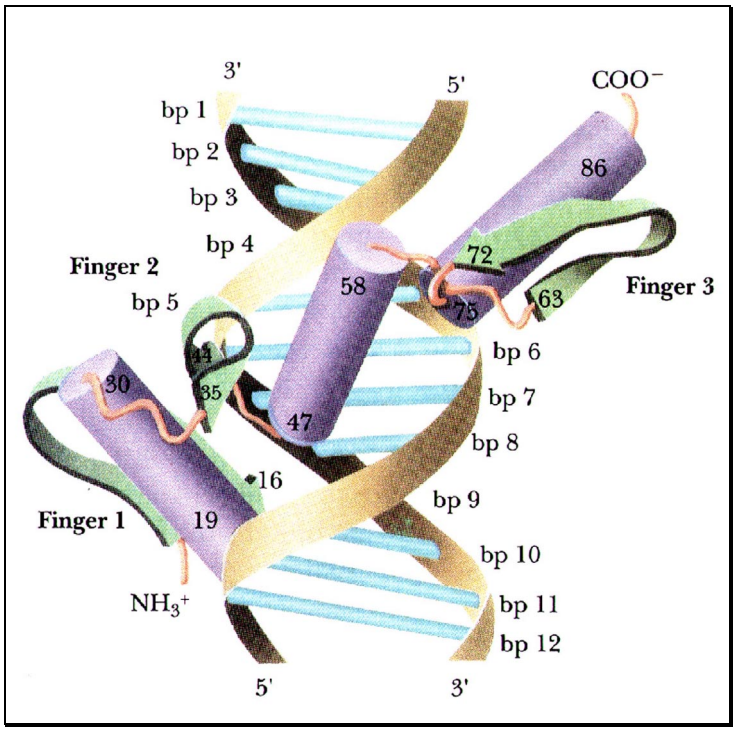


Figure 4.10. Zinc finger bound to the DNA.

IV.3.1.2. Transcription factors

Transcription factors are proteins that bind to regulatory sequences thereby modifying the rate of transcription of a gene. Some transcription factors bind directly to specific sequences in the DNA (promoters and enhancers), others bind to each other. Most of them bind both to the DNA as well as to other transcription factors. It should be noted that the transcription rate can be positively or negatively affected by the actions of transcription factors. When the transcription factor significantly decreases the transcription of a gene, it is called a repressor. If, on the other hand, the expression of a gene is upregulated, biologists speak of an activator.

The structure of a transcription factor can be divided in several domains, such as a DNA-binding domain, a transcription activation domain, and so on. The DNA binding domain determines the target genes of the transcription factor. In the case of PLAG1, the DNA binding domain consists of 7 typical zinc finger structures that recognize the GRGGC$(N)_{6-8}$RGGK motif in the DNA (Voz et al., 2000). It has been observed that genes containing several of these motifs in their promoter region are recognized by PLAG1 and their expression is initiated. The transcription activation domain is required for binding to the polymerase and actually starting the transcription process. PLAG1 contains a serine-rich COOH terminus that exhibits transactivation capacities. Transcription activation domains presumably function by direct interaction with elements of the transcription initiation complex (e.g., TFIID) or by interaction with intermediary proteins (e.g., VP16) that in turn recognize the initiation complex (Triezenberg, 1995).

IV.3.1.3. Structure of genes

Genes are segments of DNA that encode for proteins through the intermediate action of mRNA (messenger RNA). A gene usually consists of a transcribed sequence, which is converted to an mRNA, and various untranscribed regulatory sequences, consisting of the promoter region in front of the transcribed sequence, that contains transcription factor binding sites and the start site of transcription, and several upstream control elements such as enhancers. The mRNA consists of a coding sequence that is translated into a protein and several untranslated regions (UTRs). The untranscribed regulatory sequences and the sequences in the DNA that correspond to UTRs are often involved in regulation of expression, as was observed in this project for PLAG1.

IV.3.1.4. Transcription

Transcription means the assembly of ribonucleotides into a strand of mRNA (messenger RNA). The sequence of this strand of mRNA is dictated by the order of the nucleotides in the transcribed part of the gene. The transcription process is initiated by the binding of several transcription factors to regulatory sites in the DNA, usually located in the promoter region of the gene. The transcription factor proteins bind each other to form a complex that associates with an enzyme called RNA polymerase. This association enables the binding of RNA polymerase to a specific site in the promoter. In Figure 4.11, the initiation of the transcription process is shown.

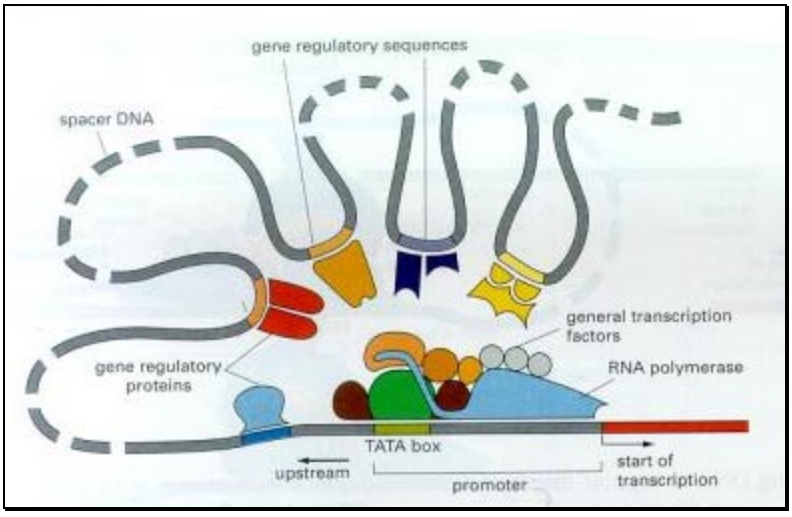


Figure 4.11. Initiation of the transcription process by the association of the complex of transcription factors (gene regulatory proteins), the RNA-polymerase, and the promoter region of a gene.

Together, the complex of transcription factors and the RNA polymerase unravel the DNA and separate both strands. Subsequently, the polymerase proceeds down on one strand while it builds up a strand of mRNA complementary to the DNA, until it reaches the terminator sequence. In this way, an mRNA is produced that is complementary to the transcribed part of the gene. Then, the mRNA transcript detaches from the RNA polymerase and the polymerase breaks it's contact with the DNA. In a later stage, the mRNA is processed, transported out of the nucleus and translated into a protein.

IV.3.1.5. Role of transcription factors in cancer

Most oncogenes activated by chromosomal translocations in leukemia's, lymphomas or solid tumors are transcription factors (Rabbitts, 1994). Since transcription factors lie at the heart of almost every fundamental developmental and homeostatic process, including DNA replication

and repair, cell growth and division, control of apoptosis and cellular differentiation, it is not surprising that inherited or acquired defects in transcription factor structure and function contribute to human carcinogenesis. Furthermore, many of these factors have been shown to have oncogenic properties when genomically altered (mutated, rearranged, amplified or deleted), transcriptionally upregulated, or post-translationally modified. Therefore, scientists have come to realize that transcription factors and the development-specific and tissue-restricted gene programs under their control represent a rich and diverse source of mechanisms, which, if disrupted, can lead to various types of malignancy. For instance, the continuous activation of PLAG1 due to promoter swapping/substitution in pleomorphic adenomas of the salivary gland results in uncontrolled transcriptional activation of downstream target genes. In the largest subgroup of these pleomorphic adenomas, exchange of a part of chromosome 8 (8q12) and a part of chromosome 3 (3p21) results in promoter swapping between the PLAG1 gene and the constitutively expressed gene for ß-catenin (Kas et al., 1997). Consequently, the transcription of PLAG1 in these adenomas is placed under the influence of a promoter that ensures continuous expression. In normal cells of the salivary gland the transcription of PLAG1 is developmentally regulated. This means that the transcription of PLAG1 in normal cells is mainly expressed in certain stages of the fetal development. In normal adult tissue, the PLAG1 is inactive. Off course the continuous expression of PLAG1 also changes the expression levels of the genes that are activated by PLAG1. For instance, if PLAG1 continuously activates the expression of genes that are involved in cell division, this could lead to uncontrolled amplification of the cell and thus the formation of a tumor. Therefore, to understand the origin of salivary gland adenomas, it is crucial to identify the genes that are targeted by PLAG1.

IV.3.2. Methodology

This project is done in cooperation with the Centre for Human Genetics in Leuven. This research group studies the functional roles of the PLAG gene family, both in tumorigenesis and in development. They have provided the microarray data that were used in the project and the required biological and medical knowledge.

IV.3.2.1. Study of the occurrences of PLAG1 binding motifs in the promoter region of a known target

During the project, the third promoter of insulin-like growth factor 2 (IGF2) was identified as a PLAG1 target (Voz et al., 2000). Information on the location of the transcription start site of IGF2, and the corresponding entry in the EMBL database, is known and can be found in the Eukaryotic Promoter Database (EPD) (Praz et al., 2002). This way the corresponding EMBL file can be retrieved and the sequence around the transcription start site can be selected with a maximal length of 1500 base pairs. Within this sequence, we searched for PLAG1-binding motifs using regular expressions. This search was done on both strands, and without allowing any overlap of the motifs. As a result of this search, IGF2 was identified as a likely PLAG1 target, which was then confirmed by several experimental procedures.

IV.3.2.2. Search for the motif in the promoters of genes upregulated by PLAG1

The occurrence of PLAG1-binding sites was studied in potential target genes that were identified by microarray analysis. A cell clone was constructed to allow regulation of the PLAG1 expression by an inducible promoter. The addition of zinc in the growth medium is necessary to induce the expression of PLAG1 in the cells. From these cells RNA was prepared, converted into cDNA and labeled, both upon (test sample) and without (reference 1) zinc induction. These cDNAs were hybridized separately on an Affymetrix chips. A second cell clone was constructed, allowing regulation of ß galactosidase by the same inducible promoter (reference 2). The labeled cDNA resulting from this experiment was mixed with the cDNA from the test sample and hybridized on a second Affymetrics chip, and on an Atlas filter (© Clontech). The genes were ranked according

to the expression ratio on the first chip. In this first experiment, both the genes upregulated by PLAG1 and other genes responding to the zinc induction, have a high ratio. In the second experiment, only the genes that are upregulated by PLAG1 result in a high ratio. The results of both experiments were combined, selecting the genes with the highest ratios in both experiments.
For the selected genes, the corresponding sequences of maximal 1,500 base pairs upstream of the translation start were retrieved from the EMBL database as before. Within these sequences, we searched for PLAG1-binding motifs using regular expressions, as described above. Based on the results of that section, it was clear that a promoter should contain multiple occurrences of the PLAG1-binding motif in order to be activated. To determine the minimum of occurrences required, the motif search results were compared to similar results obtained from a random set of 50 known promoter sequences in the EPD database. This random set was used to estimate the expected occurrence of the PLAG1-binding motif in the human genome, and more importantly the probability that the observed number of motifs n_{obs} in a sequence j of the PLAG1 dataset is generated by chance alone. This probability is calculated by the binomial

$$P(n_{obs}) = \frac{(L_j + 1 - l)!}{n_{obs}!\,(L_j + 1 - l - n_{obs})!}\, p^{n_{obs}} (1-p)^{(L_j + 1 - l - n_{obs})}$$

Here, p represents the observed frequency of the motif in the random dataset, L is the length of the selected region and l is the length of the motif. Sequences with a small P value are considered potential PLAG1 targets. Currently Northern blot experiments are performed to confirm the obtained results.

IV.3.2.3. Search for multiple occurrences of the PLAG1-binding motif in known human promoters

Based on the algorithm described in the next section, PLAG1-binding motifs can be traced on all known human promoters that are stored in the EPD database. The method is completely analogous to the one described in the previous section. However, in this case a list of about 300 sequences is obtained. The PLAG1-binding motifs are automatically obtained by a slightly adjusted version of the genome scanner, described in the next section.

IV.3.2.4. Full genome scan for multiple occurrences of the PLAG1-binding motif

Based on the results of the previous sections, it could be concluded that a minimum of 5 PLAG1 binding motifs in a relatively small region of the DNA in the proximity of the transcription start is required for activation by PLAG1. Therefore, we developed an algorithm that identifies regions in the human genome that contain multiple PLAG1-binding motifs. The sequences of the human chromosomes were retrieved from the National Centre for Biotechnology Information, Bethesda, MD. An algorithm was developed to retrieve a given chromosome sequence and to scan it for a given number of PLAG1-binding motifs. First the chromosome sequence is divided into overlapping windows (subsequences) of a given length, in most cases 500 bases pairs long. In each of the windows the number of PLAG1-binding motifs is determined using regular expressions. The sequences of all the windows are stored in a database together with the information about windows containing the required number of motifs. This way the sequences of relevant windows can be used to identify the neighboring genes that are potential targets of PLAG1. The structure of the genome scanning algorithm is outlined in Figure 4.12.

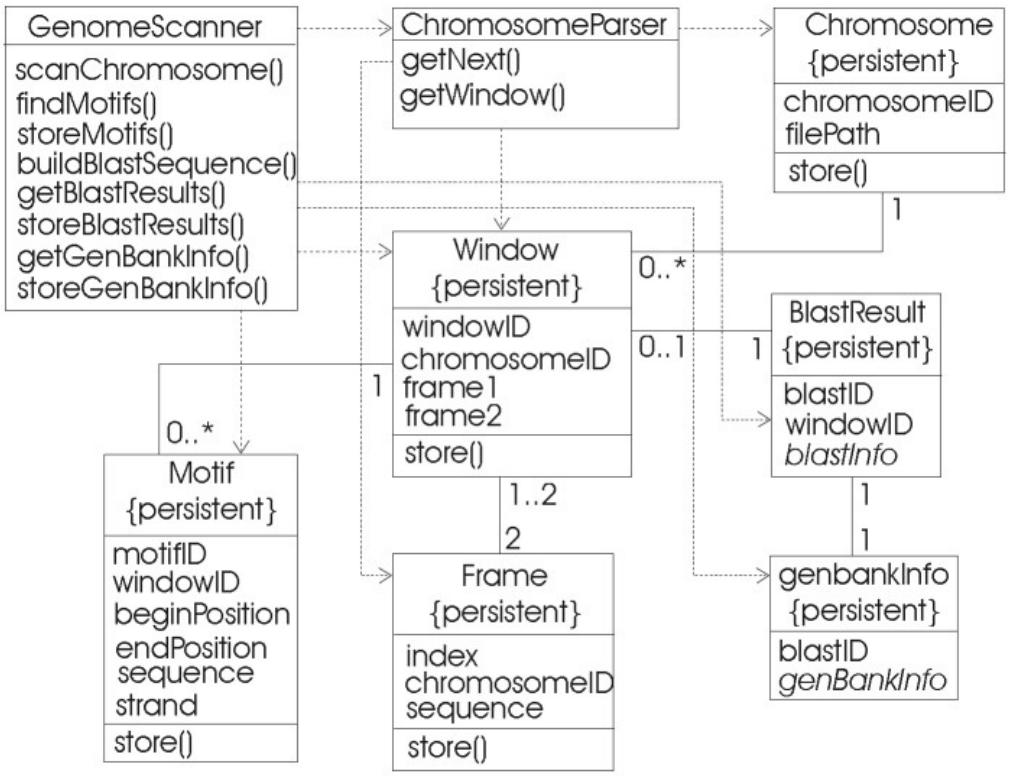


Figure 4.12. Class diagram of the genome scanner, describing the major objects in the system and the various kinds of static relationships that exist between them. Persistent objects such as 'Motif'and 'Window' are stored in the database.

IV.3.3. Results

At the start of the project, one PLAG1 target, insulin-like growth factor 2 (IGF2), was known. The Centre for Human Genetics in Leuven discovered that the third promoter of IGF2 is a target for PLAG1 regulation (Voz et al., 2000). Examination of the sequence of the third IGF2 promoter revealed that it contained a total of 8 PLAG1-binding motifs on both strands. The motifs are concentrated in a rather small region in this promoter. Sequences located further upstream from the translation start were also examined. They did, however, not contain a significant number of motifs.

IV.3.3.1. Occurrence of the motif in the promoter of genes upregulated by PLAG1

Since the promoter sequence of IGF2 contains multiple occurrences of the motif in a relatively small region located in the close proximity of the translation start, a similar strategy was used for the genes that were identified as PLAG1 targets by microarray analysis. Four potential PLAG1 target genes were identified. As expected, IGF2 is one of the predicted target genes that are upregulated by PLAG1. In this phase, all four promoters of IGF2 were analyzed and it was found that the first and fourth promoter also contain a significant number of motifs. Additionally, one of the other four potential targets, KIP2, was confirmed by Northern blotting. At this moment, the Center of Human Genetics is trying to confirm the two other potential PLAG1 targets.

To define the minimal number of motifs that a promoter of a given length should contain to be considered significant, these results were compared to the results of a similar motif search on a set of randomly selected promoters. The random dataset consisted of 50 promoter sequences from the

Eukaryotic Promoter Database. In the entire random set, merely 17 occurrences of the motif were found.

IV.3.3.2. Occurrence of the motif in known human promoters

To identify the human genes of the EPD database whose promoters contain a significant number of motifs (>=5), the sequence of the 274 known promoters was retrieved and subjected to a motif scan. The motif scanner returned 12 potential PLAG1 targets containing five to eight occurrences of the motif. Again, IGF2 pops up as a target containing 6-8 occurrences of the motif depending on the promoter. To assess whether there exists a correlation between PLAG1 activation and the chromosome on which the gene is located, for every potential target gene of the EPD database the chromosomal location was retrieved. Such a correlation could, however, not be demonstrated. Two targets are located on chromosome 11 and three on chromosome 9, of which two are located in the same region. The transcriptional activation of these genes by PLAG1 has to be experimentally confirmed by Northern blot analysis (work in progress).

Since the Center for Human Genetics also studies the tumor-inducing capacities of PLAG1 in mice, the motif scan was repeated on the known mouse promoters in the EPD database. The promoters were scanned for the same PLAG1 motif as the human promoters. None of the 196 mouse promoters contained more than 5 copies of the PLAG1 motif. There are three possible explanations for these findings. It is possible that the number of copies of the PLAG1 motif in the mouse targets is lower than in humans. On the other hand, maybe the PLAG1 motif in mouse differs slightly from the human PLAG1 motif. This can be verified by repeating the scan but each time the motif is slightly changed. Of course, it is also plausible that the EPD database simply does not contain any PLAG1 motifs.

IV.3.3.3. Full genome scan

Since we have a non-disclosure agreement with the Centre for Human Genetics, the results of the genome scan are not described in more detail. Up to now, a scan was performed for chromosome 1, 2 and 11. The chromosomes were divided into overlapping windows with a length of 500 base pairs and the windows were scanned for more than 5 occurrences of the motif. It should be noted, however, that the genome scanner generates a wealth of information and it is impossible to experimentally verify all the windows with more than 5 motifs. Therefore, methods are developed to refine the genome scanner with algorithms that are able to find additional indications of PLAG1 activation (e.g., CpG islands)

IV.3.4. Future research

Microarray experiments were performed to compare the expression levels of genes in PLAG1 expressing cells (red) with the corresponding levels in cells that express ß galactosidase instead of PLAG1 (green). Additional experiments were done for comparison of the expression levels of genes in pleomorphic adenomas (red) with the corresponding values in normal cells of the salivary gland (green). The results of all microarray analyses that were done in the course of this project should be combined to determine a list of most-likely PLAG1-targets. For the genes of this list, the promoter sequence will be retrieved and searched for PLAG1-binding motifs.

The genome scanner will be refined with algorithms for the prediction of CpG-islands, a component of the promoter that is located in the proximity of the PLAG1-binding motifs in target genes. In this way, the number of relevant windows that has to be characterized can be limited. For instance, now a scan on the rather small chromosome 11 for more than five instances of the motif returns 60 relevant windows. Tools will be developed for visualization of the results of the CpG and the motif scanner. They will incorporate information on the annotation of the sequence

that lies upstream of the relevant windows. For instance, the user can visualize information on the location of genes, transcription, and translation start.

IV.4. Motif finding in sets of co-expressed genes

The use of microarrays has opened new direction in transcript profiling research. An interesting application of this technology is the study of the transcriptional regulatory mechanism, which is responsible for the coordinated behavior of the co-expressed genes. The basic assumption states that co-expression frequently arises from transcriptional coregulation. As coregulated genes are known to share some similarities in their regulatory mechanism, possibly at transcriptional level, their promoter regions might contain some common motifs that are binding sites for transcription regulators. A sensible approach to detect these regulatory elements is to search for statistically over-represented motifs in the promoter region of such a set of co-expressed genes (Bucher, 1999; Ohler and Niemann, 2001; Zhu and Zhang, 2000). In contrast with the previous section, nothing is known about which motifs are present in the data. It is even possible that the motifs have never been described before.

Algorithms to find regulatory elements can be divided into two classes: (1) methods based on word counting (van Helden et al., 2000; Tompa, 1999; Vanet et al., 2000) and (2) methods based on probabilistic sequence models (Bailey and Elkan, 1995; Hughes et al., 2000a; Lawrence et al., 1993; Workman and Stormo, 2000). The word counting methods analyze the frequency of oligonucleotides in the upstream region and use intelligent strategies to speed up counting and to detect significantly over-represented motifs. These methods then compile a common motif by grouping similar words. Word counting methods lead to a global solution as compared to the probabilistic methods. The probabilistic methods represent the motif by a position probability matrix and they assume that the motif is hidden in a noisy background sequence. To find the parameters of this model these methods use maximum likelihood estimation in the form of Expectation Maximization (EM) and Gibbs sampling - EM is a deterministic algorithm and Gibbs Sampling is a stochastic equivalent of EM.

Within the probabilistic methods the motif is represented as a position probability matrix of dimensions 4xW, where W is the length of the motif. Each column in the matrix is discrete distribution over the four nucleotides A, C, G and T and each entry in the column gives the probability of finding a given nucleotide at that position in the motif. Figure 4.14 shows the position probability matrix that is the output of our algorithm.

We have introduced two modifications to the original Gibbs sampling algorithm by Lawrence et al., 1993 (Thijs et al., 2001). First, a probabilistic framework was used to estimate the expected number of copies of a motif in a sequence. Since both the microarray experiment and the clustering are subject to noise, only a subset of the co-expressed genes is actually coregulated. Furthermore, in higher organisms, regulatory elements can have several copies to increase the effect of the transcriptional binding factor in the transcriptional regulation. Figure 4.13 gives a schematic representation of the upstream regions of a set of co-expressed genes.

When searching for possible regulatory elements in such a set of sequences we should take into account that the motif will only appear in a subset of the original data set or could have multiple copies. We therefore want to develop an algorithm that distinguishes between the sequences in which the motif is present and those in which it is absent. We reformulated the probabilistic sequence model in such a way that the number of copies of the motif in each sequence can be estimated.

Second, we introduced the use of a higher-order background model based on a Markov chain. The use of a higher-order background model is justified by the fact that Markov models have been used extensively in the state-of-the-art gene detection software: Glimmer (Delcher et al., 1999), HMMgene (Krogh, 1997) and GeneMark.hmm (Lukashin and Borodovsky, 1998). Recently this method was also refined to interpolated Markov chains (Ohler et al., 1999). Starting from the ideas incorporated in these gene and promoter prediction algorithms, we developed a background model based on a Markov process of order m. This means that the probability of the nucleotide b at position l in the sequence depends on the m previous bases in the sequence. A transition matrix describes such a model. Important to know is that the background model can be either constructed from the original sequence data or from an independent data set. The latter approach is more sensible if the independent data set is carefully created, which means that the sequences in the training set only come from the intergenic region and thus do not overlap with coding sequences. Nevertheless the algorithm can also be used for other organisms by building the background model from the input sequences.

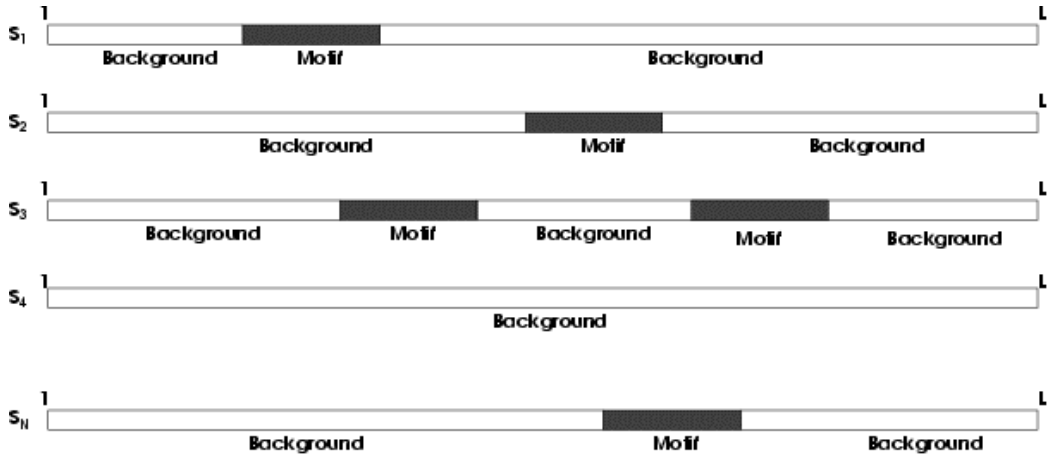


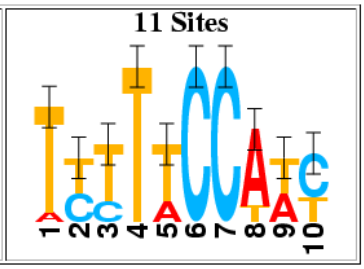
Figure 4.13. Schematic representation of a set of upstream sequences containing 0, 1, or more copies of a specific motif.

Integrating the two proposed modifications into the original Gibbs sampling algorithm for motif finding lead to our implementation, called Motif Sampler.
The input of the Motif Sampler is a set of upstream sequences. In the first step of the algorithm the higher-order background model is chosen. The background model can be pre-compiled or it can be calculated from the input sequences. The algorithm then uses this background model to compute, for each segment of length W in every sequence the probability that the segment was generated by the background model. Second, the alignment vector and the corresponding motif model are initialized. In the next step, the actual core of the sampling procedure starts. The algorithm loops over all sequences and the alignment vector for each sequence is updated. First, the motif model is calculated based on the current alignment vector. This estimated motif model is used to compute for each segment of length W in the selected sequence a weight that is the ratio of the probability that the segment is generated by this motif model divided by the probability that the segment is generated by the background model. Finally, a new alignment vector is selected by taking samples from the normalized distribution of weights over all segments in the given sequence. This procedure is repeated until the motif model has converged.

The implementation of our motif finding algorithm is part of our INCLUSive website (Thijs et al., 2002) and is accessible through a web interface: http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html. Figure 4.14 gives a screenshot of the part of the results page, which displays the position probability matrix, the corresponding logo of the motif, the motif scores and the positions of the motif instance in the sequence set.

Motif 2: TyTTTCCAwC

	1	2	3	4	5	6	7	8	9	10
A	0.0979	0.0096	0.0096	0.0096	0.2747	0.0096	0.0096	0.8049	0.4514	0.0096
C	0.0049	0.4467	0.2700	0.0049	0.0049	0.9769	0.9769	0.0049	0.0049	0.6235
G	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041
T	0.8931	0.5396	0.7164	0.9815	0.7164	0.0094	0.0094	0.1862	0.5396	0.3629



Scores

Information Content	1.1892
LogLikelihood	65.1569
Consensus Score	1.3043

Alignment:

Name	Position	Site	Prob.
Seq 1	268	TTTTTCCAAT	0.8517
Seq 2	283	TTTTACCAAC	0.9887
	635	TTCTTCCAAT	0.7876
Seq 4	512	TTTTTCCATT	0.9053
Seq 5	442	TTTTTCCAAC	0.9980
	200	TCTTTCCTTC	0.9183
Seq 6	349	TCTTACCATC	0.9994
	513	TCCTACCATC	0.9465
Seq 7	286	TCTTTCCTTC	0.9509
Seq 8	346	ATTTTCCATT	0.5764
Seq 9	229	TCCTTCCAAC	0.9806

Figure 4.14. Screenshot of the results page of the Motif Sampler.

IV.5. First steps towards integration of tools

In the previous paragraphs we discussed several tools which have been developed within our group and which are useful in the analysis of microarray data. Mostly, we have developed them as standalone tools. However, a rather interesting aspect of these tools is the possibility to integrate them in such a fashion that an in-depth exploration of the data is more feasible. This means that results from one type of analysis can be fed forward or fed back to other analysis tools. Let us give a few examples of this integration.

- Starting from the clustering of the expression profiles sets of co-expressed genes are found. We could either use the gene ontology to make a functional annotation of the genes in the cluster or we could directly search for common motifs. However, based on the functional classification of the genes more meaningful subsets of genes can be selected. Next, the upstream sequences of this selected subset of genes can be explored for common transcription factor binding sites. Using a reduced set of sequences will improve the performance of the Motif Sampler because of the expected decreased noise level in the data set. Moreover, the motifs found in a cleaner data set are assumed to be more reliable. This is a rather straightforward approach to integration. The retrieved motifs could also be used in the further analysis of the data. The top scoring motifs found in this subset are first compared to the known motifs in specialized databases.

- Another possible step is the calculation of the probability for all the other genes on the microarray that the motif is presents in their upstream region. The distribution of these scores is used to infer the significance of the motif and the specificity of the retrieved motif for the given data set. This information about motif presence can also be used in the clustering process. The original cluster is seen as a seed cluster and the cluster will be extended or shrunken based on the additional motif information. This iterative procedure can be repeated until the cluster and motifs are found in a stable configuration.
- When motifs are found in a specific cluster it might be interesting to search for other genes in the genome, which also contain the retrieved motifs in their upstream region. An example of such a genome wide scan is described in the next section.
- Since the motif finding algorithm is conceived in a Bayesian framework it will be possible to incorporate prior knowledge into the analysis. Knowledge about known transcription factor binding sites related to some genes in the cluster will improve the performance and also the interpretability of the results.
- Using the combination of the gene clusters, the functional annotation and the retrieved motifs will also be helpful in the learning of a genetic network structure. The possibility to relate a set of co-expressed genes with a known transcription factor gives valuable prior knowledge about the network structure.
- The process of integration surely requires that most steps between different analysis tools be automated. To facilitate this type of integration we are currently using and developing XML standards to pass data from one tool to another.

As an example of a first step to integration, we can refer to our web-based analysis system INCLUSive (Thijs et al., 2002). Figure 4.15 gives a high level flowchart of the process. The process starts with a text file the d microarray data. Next, adaptive quality-based clustering is used to find coherent groups of co-expressed genes. Then a semi-automated system can be used to fetch the upstream regions of the genes of interest. This system performs some time consuming searches in the genomic databases, which were typically done manually in the past. The retrieved set of sequences can then be fed into the Motif Sampler web interface to search for possible transcription factor binding sites.

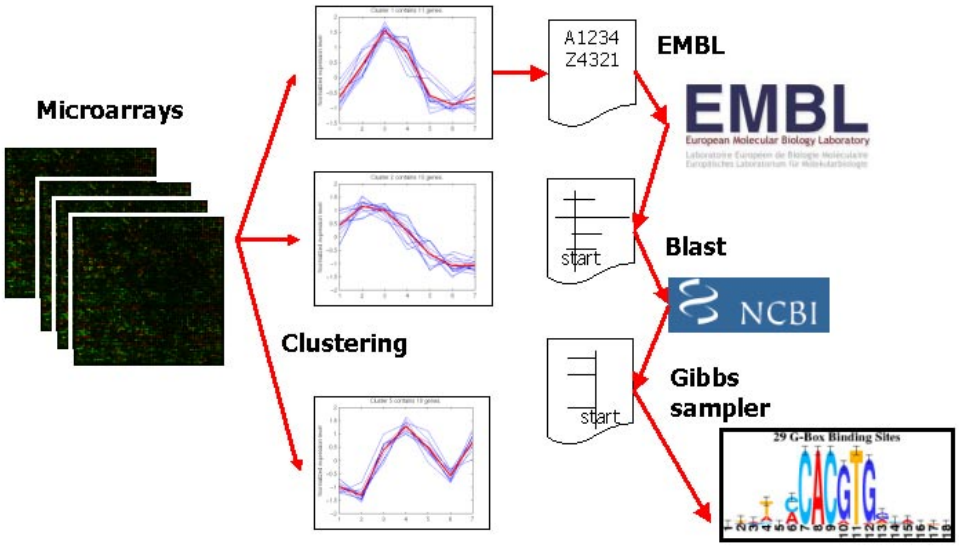


Figure 4.15. Flowchart of INCLUSive web interface. Starting from microarray data, distinct clusters are retrieved using adaptive quality-based clustering. Next the retrieval of the upstream sequences is done in a semi-automated way. Finally the sequence data can be pasted directly into the Motif Sampler.

IV.6. Conclusion

In this chapter, we studied how microarray data could be analyzed to discover elements from the pattern of gene regulation. We presented an integrated methodology that combines microarray and genome sequence data to discover which patterns control gene expression at the sequence level. The first part of the methodology is the identification of groups of genes that show similar behavior. This identification can be done by studying known groups of genes or by clustering algorithms. We presented a new clustering algorithm, called adaptive quality-based clustering, which overcomes some of the limitations of classical clustering algorithms. We demonstrated the power of this method on microarray data from the yeast cell cycle. The second part of the methodology is the identification of which short motifs in the genomic sequence are likely to control gene expression. This identification can be done either by screening the groups of differentially expressed or co-expressed genes for known motifs and detecting which motifs are common to many genes in the group or by trying to build the motif pattern from scratch using statistical methods. For scanning for a known motif, we discussed the identification of new targets of the PLAG1 transcription factor involved in benign tumors of the salivary gland. Further, we presented the method of Gibbs sampling for motif finding. We have implemented the different strategies just described in a web application called INCLUSive (Integrated Clustering and motif Sampling, http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html).

V. Towards integrative genomics

Regulatory networks are structured sets of genes and proteins that influence each other's activity. Unraveling the regulatory network helps biologists to understand the regulatory mechanisms that govern protein and gene activity. Gaining such global insight into the cellular behavior has a major impact on applied and fundamental molecular biological research. As mentioned previously, high throughput molecular biological techniques (microarray, proteomics) allow making a snapshot of the global cellular behavior and opening the door to holistic approaches. The inference of regulatory networks based on high throughput data consists of one of the central goals of molecular biology. The inference problem, however, is not straightforward. Due to its stochastic (both experimentally and biologically) and under-determined nature (mainly due to an insufficient amount of measurements), regulatory network inference poses enormous mathematical challenges. As outlined in this paragraph, it requires both the development of a robust algorithm to perform the actual inference and a knowledge system to generate the required prior knowledge.

V.1. Genetic network inference: a complex problem

In a regulatory network the connectivity between genes is hierarchically structured. In this respect genetic circuits are comparable to electric circuits. When a gene, located on top of a regulation cascade, is activated, the corresponding protein will in turn be responsible for the activation of a next set of genes (Figure 5.1). Recent molecular biological analysis techniques allow the simultaneous measurement of the expression levels of thousands of genes and proteins. It has become possible to measure the behavior of a cell induced by a regulatory network.

In many cases it is not known how the regulatory network acts, i.e., the causal relationships between the genes are unknown. Regulatory network inference is a methodology to reconstruct from experimental data (e.g., microarray data (Brown and Botstein, 1999, Blohm and Guiseppi-Elie, 2001, Southern, 2001) the underlying cellular regulatory network responsible for the observed behavior (D'haeseleer et al., 2000) (Figure 5.1). Cellular signal transduction is a multiplicative system due to the hierarchical structure of the regulatory network. This non-linearity (presence of feedback loops, synergistic effects, XOR relations) allows a cell to protect itself from extreme alterations by complex self-regulation and mutual fine-tuning of distinct regulatory cascades but poses extreme challenges for mathematics and system identification.

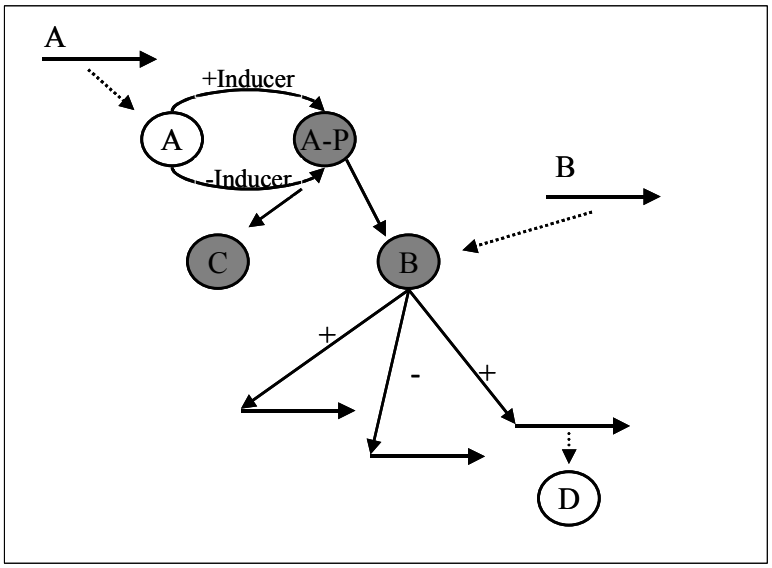


Figure 5.1. Example of a regulatory network. A key regulator (protein sensor A) is triggered by internal or external factors (inducer). By a posttranslational modification (often a phosphorylation) the key regulator is transformed into its active state and starts inducing a cascade of downstream reactions. Once activated, the key regulator on its turn regulates (activates or represses) the downstream regulators B and C. The reaction proceeds until, in a final step, DNA binding transcriptional regulator proteins (such as B) are activated. Activated or repressed transcriptional regulators influence mRNA transcription. Genes induced by the transcriptional activators are transformed into the corresponding proteins (transcription and translation process). If one of the induced genes encodes a regulator (D) the cascade continues proliferating. Bold arrows indicate Genes. Circles represent inactive proteins. Activated proteins are highlighted. P: indicates a phosphorylation. (+): activation, (-): repression. Dashed arrows represent transcription and translation processes. Other arrows represent the connectivity in the network.

Technological developments such as microarrays, proteomics, and metabolomics have opened the opportunity for massively parallel biological data acquisition. This has shifted the attention of the scientific community towards a more complex understanding of molecular biology. In addition to determining the roles of individual genes, regulatory network analysis enables us to study cells as a complex network of biochemical factors. Finding the molecular mechanisms responsible for complex diseases, drug discovery, development of sensitive diagnostics for individualized therapy, predicting the potential side-effects of a drug, and metabolic engineering are only a few examples that illustrate the enormous application potential of such holistic approach.

V.1.1. High-level methodology

Regulatory network inference implies the reconstruction of the interactions between a large number of variables (theoretically, all genes or gene products whose expression level was measured). Financial, biological and experimental restrictions, however, put a limit on the number of available measurements, which results in a seriously under-determined problem. Typically the interactions between thousands of genes need to be derived from a few hundred of experiments only. Fortunately, biology puts a number of restrictions on the number of candidate solutions. To begin with, it is known that regulatory networks are sparse. Moreover, a global network is known to consist of small modules of which only a few will be triggered by the experimental conditions tested. Based on scientific a priori knowledge (literature) or on an integrated bioinformatics analysis of the experimental data (feature extraction methods, clustering, PCA, motif finding, phylogenetic footprinting, and so on), the important units of a functional subnetwork can be identified and formally incorporated during the inference process to reduce the search space of candidate networks. Adding to the complexity of the problem is that high-throughput measurements are inherently noisy, which may lead to inconsistent observations. Moreover, the

measurements are typically incomplete in the sense that data may be missing and many variables cannot be observed (e.g., when using microarray data measurements protein-protein interactions are unobserved). These characteristics of the inference procedure and the data imply the need

- For a robust inference algorithm, able to cope with the noise in the data, with unobserved variables and creating a formal framework to incorporate prior knowledge
- For a knowledge system that retrieves information from different sources and presents it as prior information to the genetic network inference algorithm

V.2. Experimental setup

Static experiments refer to high throughput experiments that test the cellular behavior in distinct conditions or in different genetic backgrounds (e.g., the compendium experiment by Hughes et al., (Hughes et al., 2000b). If well conditioned (i.e., physiological and environmental conditions are under strict control), static experiments are very informative. In a dynamic experiment, samples are taken at regular time points from a dynamic biological system. The dynamic response of the system is caused by the action of an underlying network elicited by a perturbation and allows the system to escape from an existing steady state (non-adapted condition) to a novel steady state (adapted condition). Since the conditioning of a biological system is not straightforward, a combined experiment (dynamic profile of mutants) is most reliable. Depending on the experimental setup, regulatory networks can therefore be considered either as static or time dependent stochastic systems where the order of the underlying Markov process is most likely gene dependent. These characteristics obviously determine the type of algorithm that needs to be developed.

Though its design is generic and its content can easily be extended to other organisms (mouse, human) and other types of data analysis (protein arrays instead of the customarily used microarrays), our inference system (algorithm and knowledge system) will be developed for yeast and *Salmonella*. At this moment, ESAT-SCD is working on a project on genetic network inference in collaboration with two molecular biology labs working on yeast and *Salmonell*a respectively. These laboratories will supply the required gene expression data and expert knowledge essential for the completion of the knowledge system and the development of the genetic network inference methodology. The final goal will be the inference of genetic networks of yeast and *Salmonella* based on microarray data and prior knowledge that is supplied by the knowledge system. The results of inference procedure can be used to complement the information in the knowledge system. In this way the system can be used for fundamental research on the structure and organization of networks and pathways in yeast and *Salmonella*.

V.3. Development of a genetic network inference method in the Bayesian context

A first entity of our system for regulatory network inference is the algorithm to reconstruct the regulatory pathways from the data. In this section, we explain the advantages of Bayesian networks for this task. A Bayesian network allows both a compact representation of the joint probability distribution over a large number of variables, and an efficient way of using this representation for statistical inference. It consists of a directed acyclic graph that models the interdependencies between the variables, and a conditional probability distribution for each node with incoming edges. In the context of genetic network inference the nodes in the network represent the genes (variables). The edges correspond to the interactions. Bayesian networks are an almost natural choice to model regulatory pathways. As already pointed out, biological networks are structured hierarchically and therefore connections between genes are sparse. In a Bayesian network such sparse connections can easily be represented by the conditional independencies. Since Bayesian networks are probabilistic in nature they can capture the stochasticity (either biological or experimental) of the system. Moreover, Bayesian networks can

cope with the presence of unobserved values (hidden variables; for example, not measured protein-protein interactions). The graphical representation reflects the real biological structure and this structure can be inferred independently from the parameter estimation (maximum a posteriori, Monte-Carlo sampling).

Most important is probably the natural way by which prior information can be introduced into the model. The most important aspect of network inference is to learn as many dependencies between genes and gene products as possible from the raw expression levels of an expression profiling experiment. Given a graph it is possible to learn the probability distributions from the available data and Bayesian priors. One then searches in the space of candidate graphs for a graph that models the dependencies in the data best. The final step from a Bayesian network to a regulatory network is then a minor one.

In our current implementation, gene expression levels need to be discretized. Conditional probabilities are modeled as multinomial distributions and the conjugate Dirichlet priors are used. The identification of the highest scoring model under the Bayesian scoring metric is known to be NP-complete. Heuristic search algorithm using simulated annealing or genetic algorithms have been implemented. We do not select a single model with the highest posterior probability but compute probabilities of features of interest (edges) by averaging over the posterior model distribution. During the calculation the presence of equivalence classes is taken into account.

Further improvements to our method will include the use of

- Continuous variables to model the gene expression levels (e.g., by using Gaussian processes):
- Discretization results in loosing important information and depends on the arbitrary choice of a threshold.
- Extension of the model to dynamic systems
- Extension of BSM to handle hidden variables (unobserved data)
- Incorporation of prior knowledge. Due to the limited availability of expression data the inference problem is underdetermined. The fusion of other sources of information with the expression data can help to guide the model induction process. We will try different levels to incorporate the prior information (using non uniform prior distributions, using constrained optimization procedures etc).
- Construction of a simulated benchmark dataset to evaluate our method. This benchmark should capture as much as possible the biological complexity of real word regulatory networks

V.4. Development of an ontology-based knowledge system of yeast and *Salmonellae* for acquisition of prior knowledge

As mentioned earlier, a second major entity of our inference system is the availability of a well-designed, generic knowledge system. The knowledge system will be developed with the aim of allowing biologists to efficiently retrieve biological information on a specific organism and, most importantly, to use this knowledge for the analysis of microarray data. The knowledge system will contain a compilation of all available information together with gene expression data from relevant microarray experiments. Additionally, algorithms, tools, and newly generated information (analysis results) will be integrated in the system. The following tools will be integrated: algorithms for preprocessing, clustering (De Smet et al., 2002), classification, feature extraction and genetic network inference based on microarray data, prediction of DNA motifs in promoter sequences (Thijs et al., 2001; Thijs et al., 2002) and information retrieval from the scientific literature. The information flow between the knowledge (data)base and the tools is outlined in Figure 5.2. Our system will in the first place be developed for yeast and *Salmonella* but the design of the ontology will allow extension to other eukaryotes (human, mouse) and to the analysis of other data types (protein array data).

V.4.1. Benefits of the knowledge system

A key challenge in the bioinformatics field is the development of methodologies for inferring biological insights from large datasets such as microarray data. Therefore, the integration of tools for microarray analysis with a knowledge base has several advantages.
The knowledge base allows for sharing and *incorporation of prior knowledge* in a variety of tools. Prior knowledge can be incorporated by the use of Bayesian networks and/or decision trees (Segal et al., 2001).

Concretely, information on the function of the protein that is encoded by a gene can be used during the *clustering* process, in combination with gene expression data, for supervised clustering (Brown *et al.,* 2000) or functional annotation of unknown genes (Hvidsten et al., 2001). In supervised clustering, the expression profiles of genes that exert a function that is relevant in the context of the experiment can be used as a set of example data to classify the remainder of the data set. In the classification of yeast deletion mutants, information on the nature and the consequences of the mutations or affected pathways could be used.

As mentioned before, the incorporation of prior knowledge in Bayesian networks for genetic network inference is indispensable. Besides its role as prior knowledge, the available biological information can be used to verify the analysis results of the algorithms. Functional annotation of genes can be used to validate clusters, since co-regulated genes are thought to be often involved in the same biological process. Moreover, our in house generated microarray data can be supplemented with gene expression data from publicly available datasets (Gasch et al., 2000; Hughes et al., 2000a). The results of the Motif Sampler can be verified using information on known regulatory motifs. At the moment, part of the biological knowledge is only accessible through the *scientific literature*. Improved access to this information would therefore be very advantageous.

V.4.2. Ontologies

The knowledge model will be based on ontology in order to offer a uniform access to the knowledge implied in the system. Ontology is an abstract model of concepts and relations between concepts from a given (research) domain. Ontology in the molecular-biology field will contain concepts such as ‘gene’, ‘gene name’, ‘protein’, ‘molecular function’, ‘biological process’… Relations between these concepts can be defined. For instance, the concept ‘protein’ ‘participates in’ (relation) a certain ‘biological process’ (related concept) and ‘has’ a given ‘molecular function’. A preliminary snapshot of a small part of the yeast ontology is given in Figure 5.2.

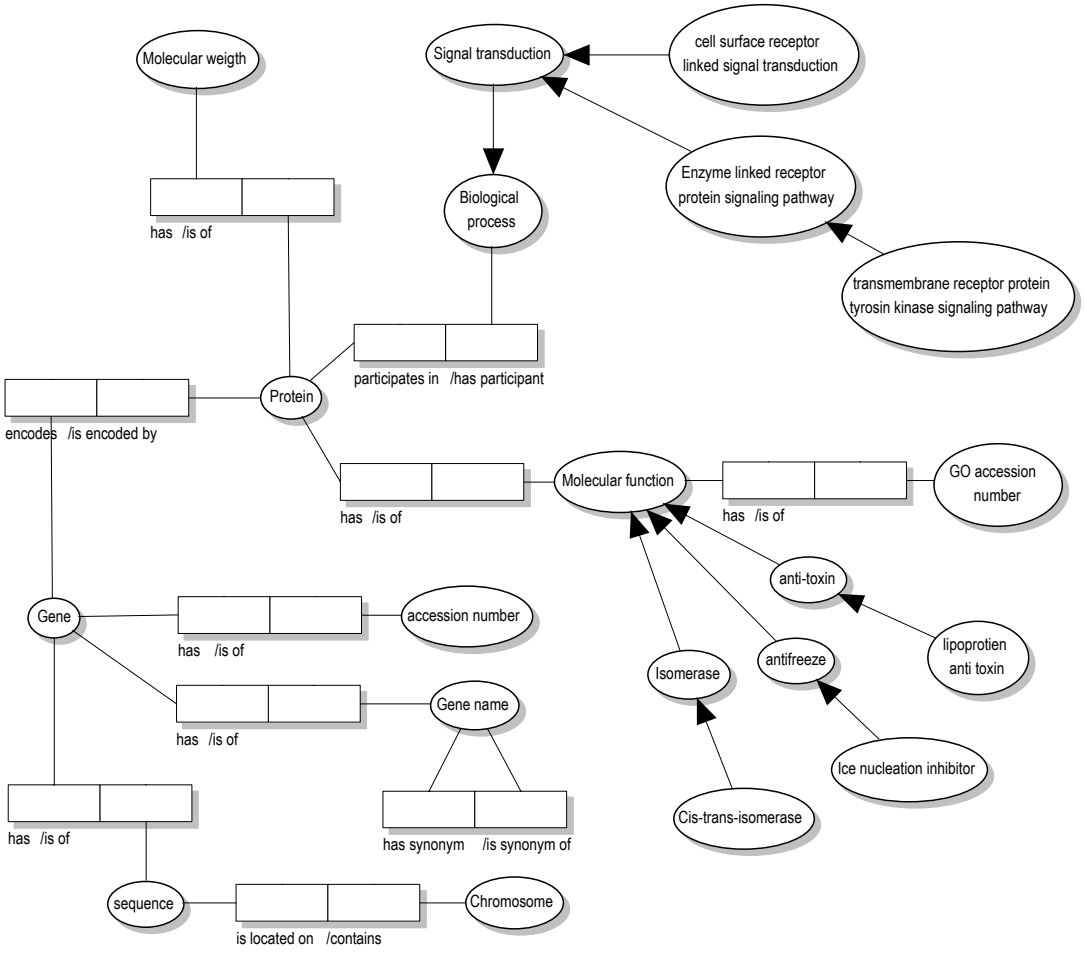


Figure 5.2. Example (preliminary) ORM-diagram of some of the core concepts and relations in the yeast ontology.

V.4.3. Benefits of ontologies

Biologists have access to multiple information resources such as public databases, sequence data, scientific literature, and public microarray datasets. A multitude of algorithms has been developed to analyze these data. Since almost every data source and algorithm uses its own data model, combining several algorithms or data sources for data analysis can be a very complex and time-consuming task. Therefore the knowledge models of different data sources need to be integrated. A classic example in the molecular biology field is the problem of genes being stored under different names in different public databases. By defining a 'is a synonym of' relation for the concept 'gene name', information retrieval from multiple data sources can be improved and simplified.

Thus, the use of ontology for the creation of a knowledge system and for the integration of different components (databases and tools) has many advantages. Because the knowledge in the system is made available in a uniform model, the ontology improves the controlled exchange of biological information among the integrated tools. A uniform representation of prior knowledge and results is also required to compare the results of algorithms with similar functionality, possibly based on different input data (e.g., clustering of text profiles vs. clustering of expression profiles) in a meaningful manner.

Usage of ontology also has advantages for the integration of system components. An *ad hoc* integration of *n* components should be avoided since in the worst case scenario this would lead to the creation of *n(n-1)* partial schemas (between each pair of components in both ways). Using a common conceptual model the integration of *n* components will only require *n* partial schemas, one between each component and the ontology.

V.4.4. Methodology

Existing ontologies (Gene Ontology (Ashburner et al., 2000), EcoCyc Mewes et al., 2002. The structure of the information in these data sources will be analyzed. Various algorithms for clustering, classification, or searching DNA motifs are already implemented at ESAT-SCD and will be integrated in the system (e.g., adaptive quality-based clustering (De Smet et al., 2002)), Gibbs sampling (Thijs et al., 2001), or support vector machines (Suykens et al., 2001)) Phylogenetic footprinting will be used for validation of DNA motifs and genetic network inference will be done using Bayesian networks. Methodologies for incorporation of biological prior knowledge in the tools will be developed based on decision trees and Bayesian networks (Segal et al., 2001). The possibilities for exchange and direct use of analysis results among algorithms will be evaluated.

Both a yeast- and a *Salmonella*-specific ontology will be created concordant with the structure of the information in existing data sources and the requirements of the algorithms using the Object-Relational Modeling (ORM) methodology. The ontology will be extended with a lexical layer, which will increase its usability by eliminating grammatical (synonyms, descriptions) and biological (heterogeneity of genetic nomenclature) differences.

The creation of the knowledge system and the integration of the algorithms will require the following steps:

1. Ontology-based specification
   The database schemes of the knowledge system will be defined according to the ontology. The ontology will also be used for the specification of the interfaces of the mediators (Wiederhold et al., 1996) that integrate the algorithms in the system.

2. Storage of the biological information
   Instances of the relations and concepts will be retrieved from public databases and locally stored. An upload tool will be created to automatically retrieve and store information from different public databases.

3. Research for the development of a ‘structured query constructor’
   The query constructor will allow the users to perform more intelligent queries based on the ontology. Navigating in the ontology can be used to generalize, expand or specialize the queries. For instance, the constructor can suggest that the query ‘DNA-binding protein’ should be limited to ‘single-stranded DNA-binding protein’, if the number of results for the first query is too high.

4. Export facilities
   The system will allow the export of information to third parties. A syntax will be developed based on RDF(S), DAML+OIL and related languages. This will not only allow for the exchange of data but also of the knowledge that is implicitly stored in these data. Additionally, the incorporation of existing ontologies, categorizations, and standards such as MIAME (Brazma et al., 2001) into our ontology, facilitates the exchange of information with third parties.

The databases and algorithms of the knowledge system can be accessed through a visual interface. The interface allows biologists to specify the research domain and the algorithms of interest and

to perform necessary analyses. New methods will be developed to compare and validate the results of various algorithms.

V.5. From local to global repositories: MIAME

In their recent paper, the MIAME workgroup underlined the importance of the use of ontologies for the description of microarray experiments. In the following a brief overview of the MIAME initiative has been described

Over the years, the sequencing of the human genome drastically changed the way biology was done, both from a fundamental and a practical stance. From a fundamental viewpoint, the human genome project (HGP) contributed in flipping the research perspective from top-down to bottom-up; It lay the seed to the exponential 'drive' currently observed in the life sciences, to an emergence of various new disciplines (most prominently bio-informatics) and to an increasing occupation of bio-technology in the business landscape. From a practical perspective, we see the biologist's lab-book gradually being replaced by a laboratory information management system and public databases take up a key role in the daily practice of science.

The recent completion of the HGP and the advent of high-throughput technologies such as microarrays, continue to amplify these trends and drive researchers in the life sciences to look at increasingly complex problems. As the problems grow more complex, so do the technologies that scrutinize them. In the previous sections and chapters, we demonstrated the numerous steps involved in a typical hybridization assay. As a result of high-throughput methodologies, researchers are literally drowning in data and are having difficulty extracting knowledge from the data due to the obstacles related to analyzing the data in the absence of common standards and (re-) implementing common functionalities to a specific architecture. Open source initiatives such as www.open-bio.org, www.openinformatics.org promote the submission of modular programs and aim to facilitate and speed up daily research at all production levels (from generating configuration files for robots to running a clustering algorithm). However, interchangeability of any biological data as well as efficient interoperability between various software components and public databases to successfully analyze them can only be guaranteed through the use of open standards. Below, we discuss how MIAME (Brazma et al., 2001) is an important step towards a standardization of microarray data and what its implications are for the research presented in this document.

As a single hybridization assay is a sensitive chain of biotechnological operations with numerous degrees of freedom regarding materials, protocols, hardware and software, it is not yet clear what *level of detail* is necessary to accompany a publication. Moreover, contrary to sequence data, expression data only makes sense in a certain context: while the sequence of a gene is uniquely determined, its transcriptome is highly dependent on the environmental conditions. Therefore an elaborate *annotation* is required to unambiguously report about an experiment.

A widely accepted standard for reporting gene expression results therefore serves a two-fold goal. Firstly, it warrants interpretability and reproducibility by providing a clear, unambiguous and complete description of the experimental conditions and protocols used in a given experiment.

Secondly, it opens the way to a technical data exchange format that encapsulates the standard. This, in turn, enables

- Creation of tools for annotating an experiment in a user-friendly manner
- Quick deployment of databases compliant to the standard
- Transparent comparison of analyses by different software packages (e.g., Image analysis)
- Transparent comparison of analyses by different technologies (Affymetrix vs cDNA)
- Easier comparison of analyses by different algorithms/methodologies
- Easier comparison of analyses by different experts

- Faster identification of benchmark datasets
- Development of mass 'in silico' experiments over numerous microarray setups

MIAME addresses these issues in an open collaborative specification for the minimum information that should be reported about a microarray experiment. Its ultimate goal is to "facilitate the establishment of databases and repositories and enable the development of data analysis tools" (Brazma et al., 2001). It is in close interaction with the MAGE-OM format and the MGED ontology, all connected through MGED – a grass roots-movement engaged to develop the standards in which we actively participate (www.mged.org). It goes without saying that various aspects of this standard directly affect our presented work around data mining, knowledge engineering and biological/medical inference.

V.6. Conclusion

We showed how Bayesian networks could be extended to the interpretation of microarray data. Bayesian networks give a structured framework to model the interaction between genes and handling the uncertainty inherent to microarray measurements; they therefore make the inference of genetic networks from microarray data possible. However, limitations in the measurements available make it essential to incorporate *a priori* knowledge. Again, Bayesian networks are the ideal platform for the integration of heterogeneous (and sometimes contradictory) sources of information. Furthermore, we describe a general framework that integrates all the methods we present and many more into a knowledge pipeline for molecular biology. There we discuss further how ontologies are central to large-scale knowledge management in genomics and we described the Minimal Information About a Microarray Experiment (MIAME) standard (Brazma et al., 2001), which our group helped develop.

VI. General conclusion

We have shown in this work how medical informatics and bioinformatics are converging into the larger field of computation, statistics, and information technology applied to biology and medicine. We call this new field Computational Biomedicine. These computational techniques provide an efficient platform to support the further integration of medicine, genetics, and molecular biology into the personalized genomic medicine of the 21st century.

As a last reminder, let us summarize the contents of this manuscript and give as references publications in which our group (ESAT-SCD (SISTA)) was involved (for a complete list of our relevant publications, please see our application file). We first showed how to build reliable tools for medical decision support using several machine-learning techniques. Logistic regression, artificial neural networks, and Bayesian networks are at the core of the tools we demonstrated. Our techniques go significantly beyond a simple proof of concept. We tackle real-world applications such as the prediction of the depth of invasion in endometrial cancer and the prediction of malignancy in ovarian tumors (Timmerman et al. 1999a, Timmerman et al., 1999b). In the framework of the International Ovarian Tumor Analysis consortium, we are currently participating to the collection of the largest database of case reports on the preoperative assessment of adnexal masses using ultrasonography. This database will serve as the starting point for the development of the next generation of intelligent tools for medical decision support. In particular, the Annotated Bayesian Network methodology appears extremely promising as Bayesian networks form a coherent framework for prediction through computational inference, for handling uncertainty, and for the incorporation of prior knowledge (Antal et al., 2001a, Antal et al., 2001b). The additional capabilities of Annotated Bayesian Networks for information management, information retrieval, and natural language decision support enhance these promises further.

In a second step, we introduced microarray technology as a major advance for molecular biology and molecular medicine. After describing the basic principles of these devices, we discussed how to manage the vast amounts of data they generate and how to transform the data from raw values into reliable measurements for further analysis. As an illustration of what then becomes possible when mining microarray data and as an extension of the philosophy advocated in the previous chapter, we developed a data-mining framework for the use of microarrays in oncology (De Smet et al., 2001). This framework involves the discovery of which genes are correlated to some clinical outcome, the prediction of an outcome on the basis of expression measurements from many genes, and the discovery of new pathology classes on the basis of the molecular fingerprint of the malignancy.

In the next step, we addressed the fundamental applications of microarrays in molecular biology, which is an even bigger field than their clinical applications. We presented an integrated methodology that combines microarray and genome sequence data to discover which patterns control gene expression at the sequence level. The first part of the methodology is the identification of groups of genes that show similar behavior. We presented a new clustering algorithm, called adaptive quality-based clustering, which overcomes some of the limitations of classical clustering algorithms (De Smet et al., 2002). We demonstrated the power of this method on microarray data from the yeast cell cycle. The second part of the methodology is the identification of which short motifs in the genomic sequence are likely to control gene expression. We studied the identification of new targets of the PLAG1 transcription factor involved in benign tumors of the salivary gland. Moreover, we presented the method of Gibbs sampling for finding potential regulatory motifs (Thijs et al., 2002). This methodology (Thijs et al. 2001) is available through our web tool INCLUSive (http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html).

Finally, we peeked into our current research topics. We showed how Bayesian networks could be extended to reconstruct networks of regulatory interactions between genes from microarray data. This task will be a major objective of molecular biologists for the coming decade. However, limitations in the measurements make it essential to incorporate *a priori* knowledge in our procedure for genetic network inference. Again, Bayesian networks are the ideal platform for the integration of heterogeneous sources of information. Additionally, we describe a strategy for the integration of multiple computational tools into a knowledge pipeline for molecular biology. In this perspective, ontologies are central to large-scale knowledge management in genomics. In particular, we described briefly the Minimal Information About a Microarray Experiment (MIAME) standard (Brazma et al., 2001) to the development of which our group participated.

Two main directions appear clearly for our future research. The first one is the further scaling of our knowledge pipeline to tackle problems in system biology. In this systems perspective, models are not limited anymore to the lowest molecular level but combine molecular processes with modeling at the cellular level, at the organ level, at the individual level, and even at the population level. The second direction is the further integration of the clinical and molecular aspects of our work by adding the genetic dimension of pathologies. In the case of oncology, it is clear that studying patterns of mutations in families or populations and correlating this information with gene expression patterns will give us a deeper view of the mechanisms of disease processes.

Let us finally conclude by saying that we are only the beginning of a long journey that will bring us well into the 21st century. We believe our work will be a modest contribution towards many revolutionary advances that will have a significant impact on society – we hope for the better.

Reference List

1. Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X., Powell,J.I., Yang,L., Marti,G.E., Moore,T., Hudson,J., Jr., Lu,L., Lewis,D.B., Tibshirani,R., Sherlock,G., Chan,W.C., Greiner,T.C., Weisenburger,D.D., Armitage,J.O., Warnke,R., and Staudt,L.M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503-511.

2. Antal P., Fannes G., De Moor,B., Vandewalle,J., Moreau,Y., and Timmerman,D. (2001a). Extended bayesian regression models: a symbiotic application of belief networks and multilayer perceptrons for the classification of ovarian tumors. Proceedings of the Eight European Conference on Artificial Intelligence in Medicine (AIME'01), Cascais, Portugal, *8*, 177-187.

3. Antal P., Meszaros,T., De Moor,B., and Dobrowiecki,T. (2001b). Annotated Bayesian Networks: a tool to integrate textual and probabilistic medical knowledge. Proceedings of the Fourteenth IEEE Symposium on Computer-Based Medical Systems, Houston, Texas, *14*, 103-108.

4. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., Harris,M.A., Hill,D.P., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J.C., Richardson,J.E., Ringwald,M., Rubin,G.M., and Sherlock,G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25-29.

5. Bailey,T.L. and Elkan,C. (1995). The value of prior knowledge in discovering motifs with MEME. Proc. Int. Conf. Intell. Syst. Mol. Biol. *3*, 21-29.

6. Baldi,P. and Long,A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics *17*, 509-519.

7. Ben Dor,A., Shamir,R., and Yakhini,Z. (1999). Clustering gene expression patterns. J. Comput. Biol. *6*, 281-297.

8. Bishop,C.M. (1995). Neural Networks for Pattern Recognition. (Oxford: Clarendon Press).

9. Blohm,D.H. and Guiseppi-Elie,A. (2001). New developments in microarray technology. Curr. Opin. Biotechnol. *12*, 41-47.

10. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C., Gaasterland,T., Glenisson,P., Holstege,F.C., Kim,I.F., Markowitz,V., Matese,J.C., Parkinson,H., Robinson,A., Sarkans,U., Schulze-Kremer,S., Stewart,J., Taylor,R., Vilo,J., and Vingron,M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet. *29*, 365-371.

11. Brown,P.O. and Botstein,D. (1999). Exploring the new world of the genome with DNA microarrays. Nat. Genet. *21*, 33-37.

12. Bucher,P. (1999). Regulatory elements and expression profiles. Curr. Opin. Struct. Biol. *9*, 400-407.

13. Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J., and Davis,R.W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell. *2*, 65-73.

14. D'haeseleer,P., Liang,S., and Somogyi,R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics *16*, 707-726.

15. De Smet,F., Marchal,K., Timmerman,D., Vergote,I., De Moor,B., and Moreau,Y. (2001). Gebruik van microroosters in de klinische oncologie. Tijdschrift voor Geneeskunde *57*, 1225-1236.

16. De Smet,F., Mathys,J., Marchal,K., Thijs,G., De Moor,B., and Moreau,Y. (2002). Adaptive quality-based clustering of gene expression profiles. Bioinformatics, in press.

17. Delcher,A.L., Harmon,D., Kasif,S., White,O., and Salzberg,S.L. (1999). Improved microbial gene identification with GLIMMER. Nucleic Acids Res. *27*, 4636-4641.

18. DeRisi,J.L., Iyer,V.R., and Brown,P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science *278*, 680-686.

19. Duggan,D.J., Bittner,M., Chen,Y., Meltzer,P., and Trent,J.M. (1999). Expression profiling using cDNA microarrays. Nat. Genet. *21*, 10-14.

20. Eisen,M.B., Spellman,P.T., Brown,P.O., and Botstein,D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U.S.A. *95*, 14863-14868.

21. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D., and Brown,P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell *11*, 4241-4257.

22. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D., and Lander,E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531-537.

23. Hastie,T., Tibshirani,R., Eisen,M.B., Alizadeh,A., Levy,R., Staudt,L., Chan,W.C., Botstein,D., and Brown,P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. *1*, RESEARCH0003.

24. Haykin S. (1994). Neural networks: a comprehensive foundation. (New York: Macmillan Publishing Company).

25. Herrero,J., Valencia,A., and Dopazo,J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics *17*, 126-136.

26. Heyer,L.J., Kruglyak,S., and Yooseph,S. (1999). Exploring expression data: identification and analysis of coexpressed genes. Genome Res. *9*, 1106-1115.

27. Hughes,J.D., Estep,P.W., Tavazoie,S., and Church,G.M. (2000a). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J. Mol. Biol. *296*, 1205-1214.

28. Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D., Kidd,M.J., King,A.M., Meyer,M.R., Slade,D., Lum,P.Y., Stepaniants,S.B., Shoemaker,D.D., Gachotte,D., Chakraburtty,K., Simon,J., Bard,M., and Friend,S.H. (2000b). Functional discovery via a compendium of expression profiles. Cell *102* , 109-126.

29. Hvidsten,T.R., Komorowski,J., Sandvik,A.K., and Laegreid,A. (2001). Predicting gene function from gene expressions and ontologies. Pac. Symp. Biocomput. *6*, 299-310.

30. Kadota,K., Miki,R., Bono,H., Shimizu,K., Okazaki,Y., and Hayashizaki,Y. (2001). Preprocessing implementation for microarray (PRIM): an efficient method for processing cDNA microarray data. Physiol Genomics *4*, 183-188.

31. Kas,K., Voz,M.L., Roijer,E., Astrom,A.K., Meyen,E., Stenman,G., and Van de Ven,W.J. (1997). Promoter swapping between the genes for a novel zinc finger protein and beta-catenin in pleiomorphic adenomas with t(3;8)(p21;q12) translocations. Nat. Genet. *15*, 170-174.

32. Kerr,M.K. and Churchill,G.A. (2001). Statistical design and the analysis of gene expression microarray data. Genet. Res. *77*, 123-128.

33. Kerr,M.K., Martin,M., and Churchill,G.A. (2000). Analysis of variance for gene expression microarray data. J. Comput. Biol. *7*, 819-837.

34. Knight,J. (2001). When the chips are down. Nature *410*, 860-861.

35. Krogh,A. (1997). Two methods for improving performance of an HMM and their application for gene finding. Proc. Int. Conf. Intell. Syst. Mol. Biol. *5*, 179-186.

36. Lander,E.S. (1999). Array of hope. Nat. Genet. *21*, 3-4.

37. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F., and Wootton,J.C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science *262*, 208-214.

38. Lerner,J.P., Timor-Tritsch,I.E., Federman,A., and Abramovich,G. (1994). Transvaginal ultrasonographic characterization of ovarian masses with an improved, weighted scoring system. Am. J. Obstet. Gynecol. *170*, 81-85.

39. Lipshutz,R.J., Fodor,S.P., Gingeras,T.R., and Lockhart,D.J. (1999). High density synthetic oligonucleotide arrays. Nat. Genet. *21*, 20-24.

40. Lukashin,A.V. and Borodovsky,M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. *26*, 1107-1115.

41. Lukashin,A.V. and Fuchs,R. (2001). Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. Bioinformatics *17*, 405-414.

42. Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S., and Weil,B. (2002). MIPS: a database for genomes and protein sequences. Nucleic Acids Res. *30*, 31-34.

43. Ohler,U., Harbeck,S., Niemann,H., Noth,E., and Reese,M.G. (1999). Interpolated markov chains for eukaryotic promoter recognition. Bioinformatics *15*, 362-369.

44. Ohler,U. and Niemann,H. (2001). Identification and analysis of eukaryotic promoters: recent computational approaches. Trends Genet. *17*, 56-60.

45. Pedersen,A.G., Baldi,P., Chauvin,Y., and Brunak,S. (1999). The biology of eukaryotic promoter prediction--a review. Comput. Chem. *23*, 191-207.

46. Perou,C.M., Sorlie,T., Eisen,M.B., van de,R.M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H., Akslen,L.A., Fluge,O., Pergamenschikov,A., Williams,C., Zhu,S.X., Lonning,P.E., Borresen-Dale,A.L., Brown,P.O., and Botstein,D. (2000). Molecular portraits of human breast tumours. Nature *406*, 747-752.

47. Pieler,T. and Bellefroid,E. (1994). Perspectives on zinc finger protein function and evolution--an update. Mol. Biol. Rep. *20*, 1-8.

48. Praz,V., Perier,R., Bonnard,C., and Bucher,P. (2002). The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. Nucleic Acids Res. *30*, 322-324.

49. Quackenbush,J. (2001). Computational analysis of microarray data. Nat. Rev. Genet. *2*, 418-427.

50. Rabbitts,T.H. (1994). Chromosomal translocations in human cancer. Nature *372*, 143-149.

51. Sager,R. (1997). Expression genetics in cancer: shifting the focus from DNA to RNA. Proc. Natl. Acad. Sci. U.S.A *94*, 952-955.

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

53. Scherf,M., Klingenhoff,A., and Werner,T. (2000). Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J. Mol. Biol. *297*, 599-606.

54. Schuchhardt,J., Beule,D., Malik,A., Wolski,E., Eickhoff,H., Lehrach,H., and Herzel,H. (2000). Normalization strategies for cDNA microarrays. Nucleic Acids Res. *28*, e47.

55. Segal,E., Taskar,B., Gasch,A., Friedman,N., and Koller,D. (2001). Rich probabilistic models for gene expression. Bioinformatics *17 Suppl. 1*, S243-S252.

56. Sherlock,G. (2000). Analysis of large-scale gene expression data. Curr. Opin. Immunol. *12*, 201-205.

57. Southern,E.M. (2001). DNA microarrays. History and overview. Methods Mol. Biol. *170*, 1-15.

58. Suykens,J.A., Vandewalle,J., and De Moor,B. (2001). Optimal control by least squares support vector machines. Neural Netw. *14*, 23-35.

59. Tailor,A., Jurkovic,D., Bourne,T.H., Collins,W.P., and Campbell,S. (1997). Sonographic prediction of malignancy in adnexal masses using multivariate logistic regression analysis. Ultrasound Obstet. Gynecol. *10*, 41-47.

60. Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S., and Golub,T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. U.S.A *96*, 2907-2912.

61. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J., and Church,G.M. (1999). Systematic determination of genetic network architecture. Nat. Genet. *22*, 281-285.

62. Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouze,P., and Moreau,Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics *17*, 1113-1122.

63. Thijs,G., Moreau,Y., De Smet,F., Mathys,J., Lescot,M., Rombauts,S., Rouze,P., De Moor,B., and Marchal,K. (2002). INCLUSive: INtegrated CLustering, Upstream Sequence retrieval and motif Sampling. Bioinformatics, in press.

64. Timmerman,D., Bourne,T.H., Tailor,A., Collins,W.P., Verrelst,H., Vandenberghe,K., and Vergote,I. (1999a). A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: the development of a new logistic regression model. Am. J. Obstet. Gynecol. *181*, 57-65.

65. Timmerman,D., Valentin,L., Bourne,T.H., Collins,W.P., Verrelst,H., and Vergote,I. (2000). Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. Ultrasound Obstet. Gynecol. *16*, 500-505.

66. Timmerman,D., Verrelst,H., Bourne,T.H., De Moor,B., Collins,W.P., Vergote,I., and Vandewalle,J. (1999b). Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses. Ultrasound Obstet. Gynecol. *13*, 17-25.

67. Tingulstad,S., Hagen,B., Skjeldestad,F.E., Onsrud,M., Kiserud,T., Halvorsen,T., and Nustad,K. (1996). Evaluation of a risk of malignancy index based on serum CA125, ultrasound findings and

menopausal status in the pre-operative diagnosis of pelvic masses. Br. J. Obstet. Gynaecol. *103*, 826-831.

68. Tompa,M. (1999). An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. Proc. Int. Conf. Intell. Syst. Mol. Biol. *7*, 262-271.

69. Tou,J.T. and Gonzalez,R.C. (1979). Pattern classification by distance functions. In Pattern recognition principles, J.T.Tou and R.C.Gonzalez, eds. (Reading, MA: Addison-Wesley), pp. 75-109.

70. Triezenberg,S.J. (1995). Structure and function of transcriptional activation domains. Curr. Opin. Genet. Dev. *5*, 190-196.

71. van Helden,J., Rios,A.F., and Collado-Vides,J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res. *28*, 1808-1818.

72. Vanet,A., Marsan,L., Labigne,A., and Sagot,M.F. (2000). Inferring regulatory elements from a whole genome. An analysis of Helicobacter pylori sigma(80) family of promoter signals. J. Mol. Biol. *297*, 335-353.

73. Voz,M.L., Agten,N.S., Van de Ven,W.J., and Kas,K. (2000). PLAG1, the main translocation target in pleomorphic adenoma of the salivary glands, is a positive regulator of IGF-II. Cancer Res. *60*, 106-113.

74. Wiederhold,G., Bilello,M., Sarathy,V., and Qian,X. (1996). A security mediator for health care information. Proc. AMIA. Annu. Fall. Symp., 120-124.

75. Wingender,E., Dietze,P., Karas,H., and Knuppel,R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res. *24*, 238-241.

76. Workman,C.T. and Stormo,G.D. (2000). ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Pac. Symp. Biocomput. *5*, 467-478.

77. Yue,H., Eastman,P.S., Wang,B.B., Minor,J., Doctolero,M.H., Nuttall,R.L., Stack,R., Becker,J.W., Montgomery,J.R., Vainer,M., and Johnston,R. (2001). An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. Nucleic Acids Res. *29*, e41.

78. Zhu,J. and Zhang,M.Q. (2000). Cluster, function and promoter: analysis of yeast expression array. Pac. Symp. Biocomput. *5*, 479-490.

79. Zweig,M.H. and Campbell,G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. *39*, 561-577.