Thèse présentée pour obtenir le titre de

Docteur de l'Université Louis Pasteur Strasbourg 1

Discipline : Sciences du Vivant
Spécialité : Bioinformatique

par

Ravi Kiran Reddy KALATHUR

Approche systématique et intégrative pour le stockage, l'analyse et la visualisation des données d'expression génique acquises par des techniques à haut débit, dans des tissus neuronaux

(An integrated systematic approach for storage, analysis and visualization of gene expression data from neuronal tissues acquired through high-throughput techniques)

Soutenue publiquement le 15 Janvier 2008 devant le jury :

Directeur de thèse	Olivier POCH, IGBMC, Illkirch
Rapporteur interne	Brigitte KIEFFER, IGBMC, Illkirch
Rapporteur externe	Christian GRIMM, UZH, Zurich
Rapporteur externe	Pascal BARBRY, IPMC, Valbonne
Examinateur	Jean-Marie WURTZ, IGBMC, Illkirch
Membre invité	Thierry LÉVEILLARD, INSERM unit 592, Paris

Dedicated to my Mother

Acknowledgements

I would like to thank the members of my jury Pascal Barbry, Christian Grimm, Jean-Marie Wurtz, Brigitte Kieffer and Thierry Léveillard for taking time out of their schedules to review my thesis.

I would also like to thank the director of the IGBMC – Dino Moras and Jean-Claude Thierry for their support.

At the outset, I would like to thank my guide and mentor Olivier Poch for giving me the opportunity to work in his laboratory over the past three years. He has been a source of inspiration and very patient with me, especially over the past few months. Working with Olivier has been a positive and enriching experience for me and will remain with me for a long time to come.

I am immensely grateful to Julie for patiently proofreading our articles and chapters of my thesis and I am indebted to her for this.

All the members of my laboratory have been a great source of strength and support to me. A special thanks to Wolfgang Raffelsberger for his constant guidance. Nicolas Wicker and I worked together towards two of my projects and he has helped me with my thesis to a great extent.

I would like to thank Nicolas and Guillaume for their extensive help with respect to the RETINOBASE project. I would also like to thank Raymond Ripp for helping me with RETINOBASE and guiding me throughout the writing of this chapter.

I would like to thank Radhouene for helping me to translate my resume in French and Anne for helping me with thesis formatting and for asking me "ça va!" every morning.

Laetitia and Naomi were the first end users of the database that I designed during my thesis; I appreciate their comments and suggestions.

My stay in the lab would not be complete if I did not mention the help from my other colleagues including Luc, Vero, Annaick, Yann, Laurent-Philippe, Yannick-Noel, Fred, Laurent, Odile, Ngoc-Hoan, David, Sophie and Emmanuel.

Once again I would like to sincerely thank Olivier Poch and his team, without their help this thesis would not have been possible.

I would like to thank Thierry Léveillard for his valuable suggestions regarding RETINOBASE.

Thanks! Serge for maintaining our servers. I would like to thank Latetia for helping me out with my documents to the university and travel arrangements.

This entire project has been supported by RETNET funded through the European Union research programme MRTN-CN-2003-504003. I would like to thank all the members of RETNET for their valuable suggestions and guidance.

On a personal front I would like to thank Raj and Ullas and their families for creating for me a home away from home and for always being there for me. All my friends here at the IGBMC who have been a stronghold and a great support to me, I am glad I have friends like you. Murugan, Beena, Jai, Amit and Harshal thanks for being around!

Last but not least my heartfelt gratitude goes out to my dear parents, my brother and sister-in-law for always being there for me. This would not have been possible without you, and so I dedicate this work to you.

List of abbreviations

Abbreviations (Computer and Statistics)

AIC: Akaike Information Criterion

BIC: Bayesian Information Criterion

CEM: Competitive Expectation Maximisation

CGED: Cancer Gene Expression Database

CSS: Cascading Style Sheets

DBMS: DataBase Management System

EM: Expectation Maximisation

FTP: File Transfer Protocol

HTML: Hypertext Markup Language

LDAP (netscape): Lightweight Directory Access Protocol

MLA: Maximum Likelihood Approximation

ODBMS: Object Database Management System

PHP: Hypertext Preprocessor

RDBMS: Relational Database Management System

SQL: Structured Query Language

Abbreviations (Biology and Bioinformatics)

CEL: Contains information about each probe on the chip is extracted from the image data by the Affymetrix image analysis software.

dChip: DNA-Chip Analyzer

DDBJ: DNA Data Bank of Japan

DNA: Deoxyribonucleic acid

EMBL: The European Molecular Biology Laboratory

EST: Expressed Sequence Tag

FASABI: Functional And Statistical of Biological Data

GEO: Gene Expression Omnibus

GO: Gene Ontology

GOLD: Genomes OnLine Database

KEGG: Kyoto Encyclopedia of Genes and Genomes

MAS 5.0: Affymetix Microarray Suite 5.0

MIAME: Minimum Information About a Microarray Experiment

OMIM: Online Mendelian Inheritance in Man

PCR: Polymerase Chain Reaction

RD: Retinal Disease

RETNET: European Retinal Research Training Network

RISC: RNA-induced silencing complex

RMA: Robust Multi-array Analysis

RNA: Ribonucleic acid

SAGE: Seiral Analysis of Gene Expression

SIEGE: Smoke Induced Epithelial Gene Expression

SNP: Single Nucleotide polymorphism

SPR/SPRR: Small Proline-Rich Proteins

SR: Serin/Arginine

UTR: UnTranslated Region

List of contents

List of Figures

List of Tables

Chapter 1. Avant-Propos

The work presented in this manuscript concerns different aspects of gene expression data analysis, encompassing statistical methods and storage and visualization systems used to exploit and mine pertinent information from large volumes of data. Overall, I had the opportunity during my thesis to work on these various aspects firstly, by contributing to the tests through the design of biological applications for new clustering and meta-analysis approaches developed in our laboratory and secondly, by the development of RETINOBASE, a relational database for storage and efficient querying of transcriptomic data which represents my major project.

The first project is related to the development of a new method, called the Maximum Likelihood Approximation (MLA) and developed by Nicolas Wicker, for the estimation of the parameters of a Dirichlet distribution. Apart from the uniform distribution, the Dirichlet distributions are the simplest distributions used to efficiently analyze data with proportions. In addition to numerous applications in diverse scientific domains, Dirichlet distributions are now being used in a growing number of applications in biology. To quote an example, Dirichlet distributions have been used to represent proportions of amino acids when modeling sequences with hidden Markov models (Sjolander *et al.*, 1996). By comparing the functional enrichment of protein clusters, obtained either by the MLA or moments methods in a mixture model background, based on their amino acid composition, we could verify that the MLA method represents a robust and rapid method suitable for the analysis of the large volumes of data generated by the post-genomic era.

The second project is related to gene expression data analysis, through the design of biological applications to test a new class of methods called MDF (Multi-Dimensional Fitting) that have been developed in our laboratory by Nicolas Wicker, in collaboration with Claude Berge and Dr. Nicolas Froloff. This method transforms datasets of one target matrix to fit distances computed on a reference matrix. Taking advantage of RETINOBASE and of the various tools and strategies implemented for transcriptomics data treatment and analysis, we applied the MDF method to investigate it's robustness in comparative meta-analysis of large-scale transcriptomic experiments. The *in silico* functional characterisation of the gene sets exhibiting statistically significant expression behaviour after the MDF transformation revealed that this new class of methods might represent a very powerful and efficient approach for modern high throughput data analysis.

The main project of my thesis is related to the development of a database that stores transcriptomic data, offers various specialized querying and display options as well as provides a platform to compare various normalization and clustering algorithms. The database was developed in the context of the "European Retinal Research Training Network" (RETNET) of the 6th Framework Programme (FP6), involving a consortium of 9 different laboratories dedicated to retinal research. RETINOBASE has been developed in close collaboration with various laboratories in the consortium to store gene expression data from retina. Initially, RETINOBASE was started with private datasets from experiments performed in the laboratories of Dr. Thierry Leveillard and Dr. Peter Humphries. Over time, the database has grown in volume and currently RETINOBASE harbors retinal gene expression data from 30 different experiments both public and propriety, performed on 6 different organisms. Most of the datasets in RETINOBASE are analyzed using a variety of background correcting and normalization software such as RMA (Robust Multi-array Analysis), dChip (DNA-Chip Analyzer) and MAS 5.0 (Affymetrix Microarray Suite 5) and data obtained after normalization are clustered using K-means, Mixture models algorithm. The motivation for maintaining distinct normalization and clustering methods has strong implications in the database structure and in the querying and visualization developments, but this enabled RETINOBASE to serve as a platform to compare these methods. For example, the gene expression information, especially the distinct cluster information available in RETINOBASE, has been successfully exploited in identifying more than a thousand genes for RetChip, an oligonucelotide microarray developed in close collaboration with retinal experts and dedicated to the study of retinal development and degeneration.

Overview of introductory sections

The rapid improvement and development of various technologies in fields of biology such as genomics, transcriptomics and proteomics has resulted in a huge explosion of data. This post-genomic biology results from the synergy between the developments in informatics, biology and biotechnology that will be quickly presented in chapter 2.

In this context, the various database systems that have been implemented or adapted to store, as well as to give access to and to query efficiently the different levels of biological data will be presented in chapter 3, in addition to a rapid presentation of the major biological databases related to sequence, function, gene expression and pathway data.

Transcriptomics (detailed in chapter 4) plays a major role in understanding the expression patterns of various genes in a cell, tissue or organism. After a rapid description of the major mechanisms involved in precise control of gene expression, chapter 4 presents the various techniques available to obtain the transcriptomic data. Chapter 5 deals with the different aspects of microarray data analysis starting from experimental design, description of the different algorithms available for data normalization and clustering, and finishing with an emphasis on meta-analysis of gene expression data.

Finally, a general description of the retina, different retinal diseases and various models available to study the pathogenesis of these diseases is discussed in chapter 6.

Introduction

Chapter 2. Biology and Bioinformatics

Biology has been transformed by the availability of numerous complete genome sequences for a variety of organisms ranging from viruses or bacteria up to plants and animals, such as mouse or human. Embedded within this as-yet poorly understood code are the genetic instructions for the entire repertoire of molecular functions and cellular components, knowledge of which is needed to unravel the complexities of biological systems. Bioinformatics is expected to play a crucial role in the systematic interpretation of genome information and the associated data from other high-throughput experimental techniques, such as structural genomics, transcriptomics and proteomics. The results of such large-scale analyses will have widespread implications for fundamental research.

In this chapter, I will quickly describe some major steps involved in the central dogma of molecular biology and also provide a time line of major events in the fields of informatics, molecular biology and bioinformatics. This is followed by a statistics on mouse and human genome sequence. Finally, I will end this chapter with a focus on the interdependency between computational and experimental techniques that play a critical role in the new systems level biology.

2.1 Central dogma of molecular biology

The modern era of molecular biology began in 1953 when James D. Watson and Francis H.C. Crick proposed the double helical structure of DNA (Deoxyribonucleic acid) (Watson and Crick, 1953) with the help of X-ray diffraction studies of DNA performed by Rosalind Franklin. In 1958, Francis H.C. Crick coined the term “the central dogma of molecular Biology” (Crick, 1958) (Figure 1) to specify the flow of genetic information that plays a central role in the development and organization of living organisms.

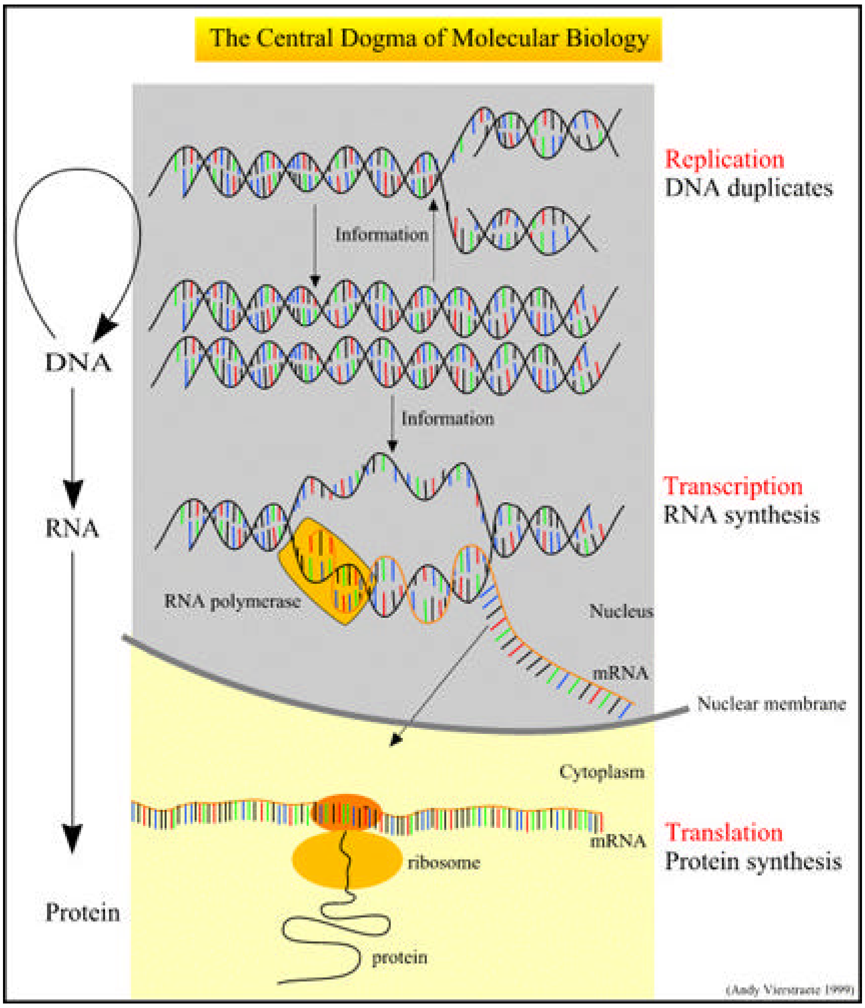


Figure 1. The Central Dogma of Molecular Biology.
Flow of genetic information, The central dogma forms the backbone of molecular biology and includes three major processes involved in the flow of genetic information namely: Replication: the process of copying a double stranded DNA molecule. Transcription: the process through which a DNA sequence is enzymatically copied by an RNA polymerase to produce a complementary copy of RNA. Translation: the information in mRNA (messengerRNA) is decoded by ribosomes to produce a specific polypeptide in accordance with genetic code. Figure adapted from MITOPENCOURSEWARE (http://ocw.mit.edu).

This concept of a flow of genetic information, together with reverse transcription that was discovered later, in combination with the development of novel techniques not only in biology but also in informatics, resulted in the new field of bioinformatics.

2.2 Timeline of major events in informatics, molecular biology and bioinformatics

This section of the chapter details various events in biology and informatics along with those that eventually led to the integration of these two fields resulting in the emergence of an inter disciplinary field called Bioinformatics. I would like to emphasize on a few milestones in relation to the development of this integration through this chapter.

The history of biology dates back to the early 19 century with the convergence of various scientific disciplines. Later, towards a better understanding of biological processes scientists took an interest in what came to be termed “molecular biology”. The most important revelation to science today, is probably the description of the structure of DNA in 1953 by Watson and Crick. Watson and Crick's model laid out the "Central Dogma", which foretold the relationship between DNA, RNA, and proteins. A critical confirmation of the replication mechanism that was implied by the double-helical structure followed in 1958 from Meselson-Stahl experiment. These findings represent the birth of molecular biology.

It was also around the same time that Howard Temin and David Baltimore independently isolated reverse transcriptase, an enzyme that can synthesize DNA from RNA (1970). In 1977, Allan Maxam, Walter Gilbert and Fredrick Sanger developed different methods for sequencing DNA and Kary Mullis (1983) described the PCR reaction. The automation of sequencing had lead to a drastic rise in sequencing projects which generated loads of data which need to be stored and queried efficiently to obtain useful and exploitable information from post-genomic data. In the 80’s, computational methods were developed to organise and analyze the data stored in the first biological databases and the field of bioinformatics took shape as an independent research discipline. The central concept of a database is that of a collection of records. One of the most common kind of databases management systems in use today - Relational Database Management System, that conform to the relational model and refers to a database's data and schema, was described by Codd Ef. This model represents relationships by the use of values common to more than one table.

Considering needs of the scientific community, European Molecular Biology Laboratory (EMBL) in 1980 created the first nucleic acid sequence database. Subsequently, a number of organizations developed sequence databases like Genebank (1982), DNA Data Bank of Japan (DDBJ, 1986) and National Centre for Biotechnology Information (NCBI, 1988).

More recently, Schena and collaborators in 1995 described the quantitative monitoring of gene expression patterns with a complementary DNA microarray and since then the microarray field has expanded to include more than 20,000 scientific publications. Hardly about a year after this, in 1996, Affymetrix marketed its first DNA chip! With the emergence of high throughput techniques the need for information management systems also rose, to collect and store as well as to curate and cross validate heterogeneous information, allowing efficient data retrieval and exploitation. These developments are opening up the possibility of new large scale studies of complex biological systems.

Year	**Person/Company**	**Discovery/Invention**
1905-1908	William Bateson and R. C. Punnett	The first time gene regulation was demonstrated
1950	Erwin Chargaff	Found that ratios of adenine to thymine and cytosine to guanine in DNA are always about the same. "Chargaff's Rule"
	Alfred Day Hershey and Maratha Chase	Proved on the basis of bacteriophage, that DNA is the genetic material
1952	Rosalind Franklin and Maurice Wilkins	Performed X-ray crystallography studies of DNA, providing crucial information that led to the elucidation of the structure of DNA
1953	James Watson and Francis Crick	Proposed the double-stranded, helical, complementary, anti-parallel model of DNA
	IBM 650	First commercial computer
1955	Fredrick Sanger	Determined the complete sequence of a protein, bovine insulin
	Arthur Kornberg	Discovered and isolated DNA polymerase from E.coli bacteria
1956	Christian B Anfisen	3-D conformation of proteins is specified by their amino acid sequence
1958	Francis Crick and George Gamov	Enunciated the "central dogma" of molecular biology: Genetic information flows in one direction from DNA to mRNA to protein
1961	Sidney Bernner and Francis Jacob	Identification of messenger RNA (mRNA)
	Marshall Nirenberg, Har Gobind Khorana and Servo Ochoa	Cracked the "Genetic code": a sequence of three nucleotide bases (codon) determine each of amino acids
1965	Margaret O. Dayhoff	The first atlas of protein sequence and structure, which contained sequence information on 65 proteins
1969	ARPANET	Computers linking between several universities

1970	Howard Temin and David Baltimore	Independently isolated reverse transcriptase, an enzyme that can synthesize DNA from RNA
	Codd E.F	Proposed relational model for the database
1972	Paul Berg	First successful DNA cloning experiments
1974	Vinton Gray Cerf and Bob Kahn	Development of the concept of connecting networks of computer into an “internet” and develop the Transmission Control Protocol (TCP)
	Chou PY and Fasman GD	Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins
	Charles Goldfarb	Invents SGML (Standardized General Markup Language)
1975	Edwin Southern	Published experimental details for the Southern Blot technique to identify DNA fragments
1976	Bell labs	The Unix-to-Unix protocol (UCCP)
	Dr. R. M. Metacalfe	Develops Ethernet, will allowed coaxial cable to move data faster
1977	Allan Maxam, Walter Gilbert and Fredrick Sanger	Independently developed different methods for sequencing DNA
1978	Fredrick Sanger	The first complete genome sequence for virus (pi-x 174) (5386 bp)
1980	EMBL	Creation of the first European nucleic acid sequence database
1981	Temple Smith and Michael S. Waterman	Algorithm for local optimal sequence alignment
	IBM	IBM introduces its personal computer into the market
1982	Genbank	Creation of American database of nucleic acid sequences
1983	Kary Mullis	Invention of the polymerase chain reaction (PCR)
1984	Gouy	ACNUC: software of inquiry of sequences databases
1985	Lipman and Pearson	FASTA
1986	Swiss-Prot	Creation of the protein sequences database
	DDBJ	Creation of Japanese, nucleic acid sequence database
	Thomas Roderick	Invented the “genomic” term
1987	Applied Biosystems	Marketing first automated sequencer
	David Burke	Created expression vector “Yeast artificial chromosomes”, **YACS**
	Mc Kusick	First genetic map of the human genome
	DNA chip	Creation of technology
1988	Higgins DG	CLUSTAL: software for multiple alignment of sequences
	Tagle DA	Phylogenetic footprinting
	Edgar Wingender	TRANSFAC database
	NCBI	National Centre for Biotechnology Information (NCBI) founded at NIH/NLM
1990	Altschul S. F	BLAST: Fast sequence similarity searching tool

	Tim Berners-Lee	Publication of first HTML document
	Francis S Collins	Development of positional cloning
	HGP	Human Genome project launched: estimated cost $13 billion
1991	Mark D. Adams	First high-throughput sequencing of cDNA (EST): a method for identifying active genes
	Roberts L	GRAIL: gene localization program
	Linus Torvalds	Announces a Unix like operating system, later termed as LINUX
1992	Mel Simon	Introducing the use of BACs for cloning
1993	Cohen D	First physical map of the human genome
	Boguski MG	DbEST: international database for EST
	Etzold T	SRS: program for database request
1994	Julie D Thompson	ClustalW, multiple sequence alignment
1995	Schena M	First paper on microarrays (Schena *et al*., 1995)
	Fleischmann RD	Sequencing of the first living organism, the bacteria *Haemophilis influenzae* (1.8 Mbp)
1996	Walsh S	Sequencing of the first eukaryotic genome *Saccharomyces cerevisiae* (12.1Mbp)
	Affymetrix	Marketing of first DNA chip
1997	Altschul S.F	Gapped Blast
	Blattner FR	Escherichia coli sequencing (4.7 Mbp)
	Chirs Burge	GenScan program: complete gene structures prediction in human genomic DNA
	Ross Ihaka	R statistical system started
1998		Sequencing of first multicellular organism, *Caenorhabditis elegans* (97 Mbp)
1999	Dunham I	Sequencing of human 22 chromosome
2000	Carina Dennis	*Arabidopsis thaliana* genome sequencing (100 Mbp)
	Adams M. D	*Drosophila melanogaster* genome sequencing (180 Mbp)
2001	Lander E.S	Preliminary sequence of human genome (3 Gbp) by HGP
	Craig Venter	Preliminary sequence of human genome (3 Gbp) by HGP by Celera genomics
	Li. C and Wong H.W	DNA –Chip Analyser (dChip) for probe and high level analysis of gene expression data
2002	Waterston	Preliminary sequence of mouse genome (2.5 Gbp)
	Suzuki Y	Database of Transcriptional start sites, DBTSS
	Robert C. Gentleman	First Bioconductor software release
2003	Karolchik D	UCSC Genome Browser database
2004	Gibbs R.A	Sequence of Whole rat genome

	Hiller L.W	Sequence of whole genome of *Gallus gallus*
	Irizarry RA	RMA method for Affymetrix GeneChip array analysis
	Stalker J	ENSEMBL genome browser
	HGP	Human gene counts estimates changed to 20,000 to 25,000
	ENCODE	Identifying all functional elements in the human genome sequence initiated
Soon ……..		More than 600 eukaryotic genomes are currently being sequence

Table 1. Chronology of key events in biology, informatics and bioinformatics.
(adapted from http://biocc.kobic.re.kr/Bioinformatics/history_of_bioinformatics.html
and http://www.genome-informatics.net/webportal/background/timeline.html)
The colours **black**, **blue** and **red** describes events in biology, informatics and bioinformatics respectively. Pertinent advances particularly concerning databases or gene expression are depicted in green and some of them are described in more detail in the manuscript.

2.3 Genome sequencing

Central to the emergence of all post-genomic biotechnologies and analysis is the availability of complete genome information. Sequencing of many complete genomes resulted from the invention of sequencing technology separately by Sanger (dideoxy method) (Sanger *et al.*, 1977) and Maxam and Gilbert (Chemical cleavage method) (Maxam and Gilbert, 1977) and the later development of automated versions. The first free-living organism to be sequenced was *Haemophilus influenzae* (1.8Mb) in 1995, and since then, genomes have been sequenced at an increasingly rapid pace. In early 2001, the International Human Genome Sequencing Consortium reported a first draft sequence covering about 90% of the euchromatic human genome, with about 35% in finished form. Since then, progress towards a complete human sequence has proceeded swiftly, with approximately 98% of the genome now available in the draft form and about 95% in the finished form. Many of the model organisms that have been completely sequenced since then, occupy strategic positions in the tree of life and provide important information for evolutionary studies or for commercial uses (Delsuc *et al.*, 2005). Comparative analyses of the numerous genome sequences now available, has become a powerful approach, whose goal is to discriminate conserved from divergent and functional from non-functional DNA. It is now providing insights into the DNA sequences encoding proteins and RNAs responsible for conserved molecular and cellular functions, as well as the DNA sequences controlling the expression of genes. This approach is also contributing to the annotation of newly sequenced genomes and the identification of functional DNA segments, such as coding exons, noncoding RNAs, and some gene regulatory regions. In addition, DNA sequencing has many applications in human studies, such as the search for genetic variations in dedicated populations and/or for mutations that may play a role in the development or progression of a disease. The disease-causing change may be as small as the substitution, deletion, or addition/deletion of a single base pair or as large as the addition/deletion of thousands of bases.

Currently (September 2007), Genomes OnLine Database (GOLD: http://www.genomesonline.org/) contains about 634 complete genomes that are published (Figure 2) and a further 771 eukaryotic, 1290 bacterial and 60 archeal genomes are being sequenced.

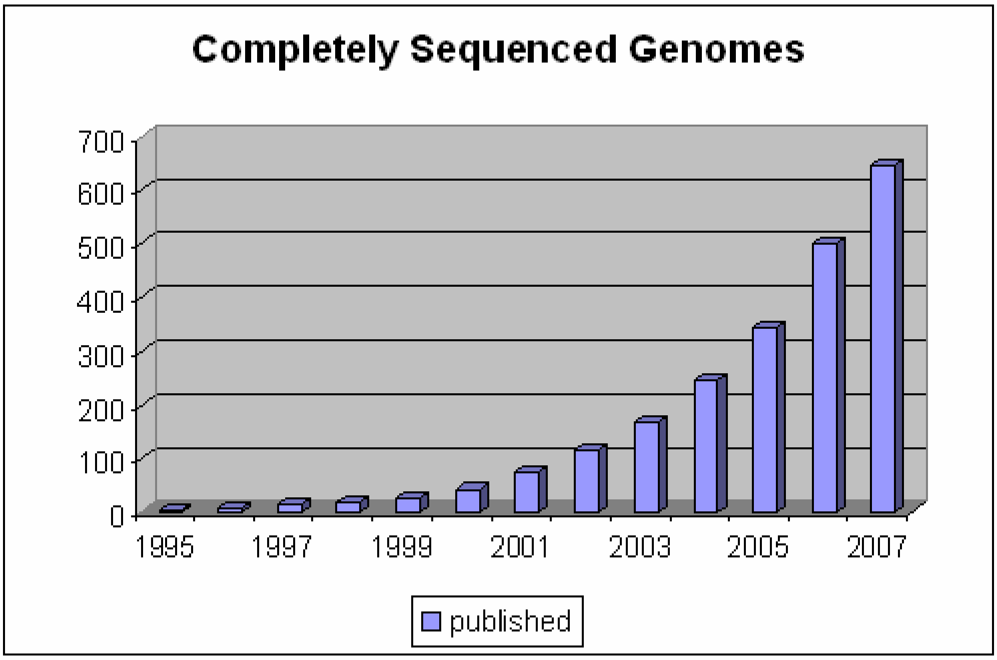


Figure 2. Illustration of the rapid growth in the number of seque nced genomes. The X-axis indicates the year and theY-axis represents the number of genomes completely sequenced. Figure adapted from **http://www.genomesonline.org/**

2.4 Interdependency between computational and experimental techniques

Fields such as genomics, transcriptomics or proteomics are built on the synergism between computational and experimental techniques. This type of synergism is especially important, not only in accomplishing goals such as identifying novel molecular mechanisms or defining signalling cascades and molecular interactions, but will also contribute to speed-up experimental studies of complex eukaryotic systems by providing a directed approach to solving complex problems. Computational techniques, by definition, are predictive and vary in performance quality. Experimental results provide a spectrum of information, ranging from implied functional relevance to validation of protein identity. Figure 3 illustrates this interplay between computational and experimental strategies that is central in modern high throughput biology.

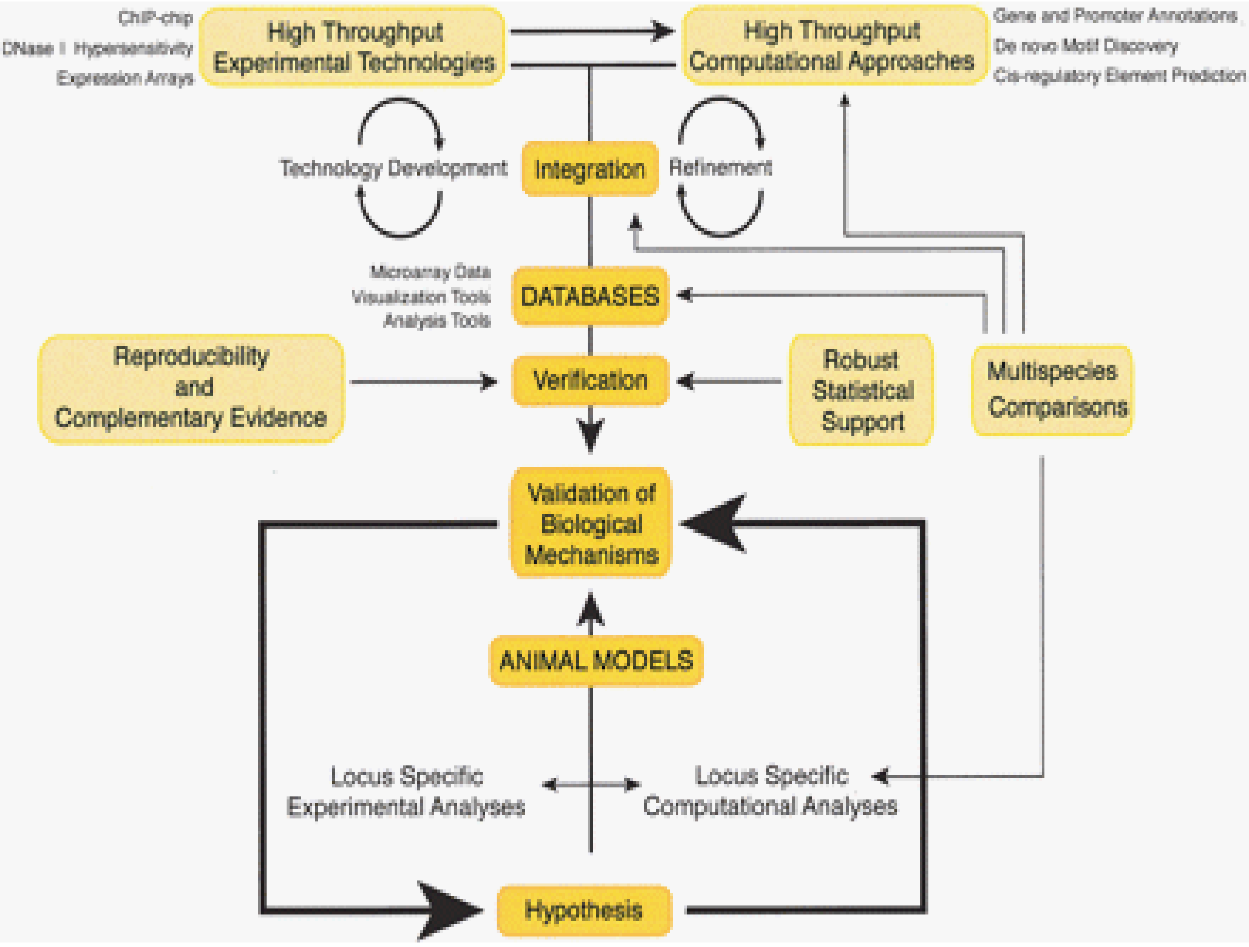

ChIP-chip
DNase I Hypersensitivity
Expression Arrays
High Throughput Experimental Technologies
High Throughput Computational Approaches
Gene and Promoter Annotations
De novo Motif Discovery
Cis-regulatory Element Prediction
Technology Development
Integration
Refinement
Microarray Data
Visualization Tools
Analysis Tools
DATABASES
Reproducibility and Complementary Evidence
Verification
Robust Statistical Support
Multispecies Comparisons
Validation of Biological Mechanisms
ANIMAL MODELS
Locus Specific Experimental Analyses
Locus Specific Computational Analyses
Hypothesis

Figure 3. The interplay and co-dependency of experimental and computational approaches. The schema illustrates the interplay between computational and experimental strategies. All indicated pathways lead to the ultimate goal of validation of a biological mechanism. Figure adapted from (Elnitski *et al.*, 2006) .

Chapter 3. Biological Databases

In this chapter, I introduce two different types of database management systems the Object Database Management System (ODBMS) and Relational Database Mangament System) with more emphasis on RDBMS. Then, the roles of biological databases, specific features, and challenges will be quickly discussed. I briefly describe (architecture, functionalities and uses) of a number of important biological databases (sequence, structure, ontology databases) and querying systems like Entrez and SRS. Finally, microarray databases and their functionalities and architecture are described.

3.1 General introduction to databases

A database is a structured collection of records or data that is designed to offer an organized mechanism for storing, managing and retrieving information. The computer program used to manage and query a database is known as a DataBase Management System (DBMS). Typically, for a given database, there is a structural description of the type of facts held in that database: this description is known as a schema. The schema describes the objects (tables) that are represented in the database, and the relationships among them. There are a number of different ways of organizing a schema. The model in most common use today is the relational model.

3.1.1 Object Database Management System (ODBMS)

An object database management system (ODBMS) is a database management system (DBMS) that supports the modeling and creation of data as objects (object is a self-contained entity that consists of both data and a section of software programme that performs specific task).

There is currently no widely agreed-upon standard for what constitutes an ODBMS and no standard query language to ODBMS equivalent to what SQL is to RDBMS to query the database. ODBMS were originally thought of to replace RDBMS because of their agreeability with object-oriented programming languages like C++ and JAVA. However, high switching cost, the inclusion of object-oriented features in RDBMS, has made RDBMS successful.

3.1.2 Relational Database Management System (RDBMS)

A relational database (Figure 4) is a collection of data items organized as a set of formally-described tables from which data can be accessed or reassembled in many different ways without having to reorganize the database tables. The relational database was invented by E. F. Codd at IBM in 1970. A relational database is a set of tables containing data fitted into predefined categories. Each table (which is sometimes called a relation) contains one or more data categories in columns.

Properties of Relational Tables:

- Each row is unique
- Each column has a unique name
- The sequence of rows is insignificant
- The sequence of columns is insignificant
- Column values are of the same kind

Certain fields may be designated as a key, which means that searches for specific values of that field will use indexing to speed them up. Where fields in two different tables take values from the same set, a join operation can be performed to select related records in the two tables by matching values in those fields. Often, but not always, the fields will have the same name in both tables. Since the relationships are only specified at retrieval time, relational databases are classed as dynamic database management systems.

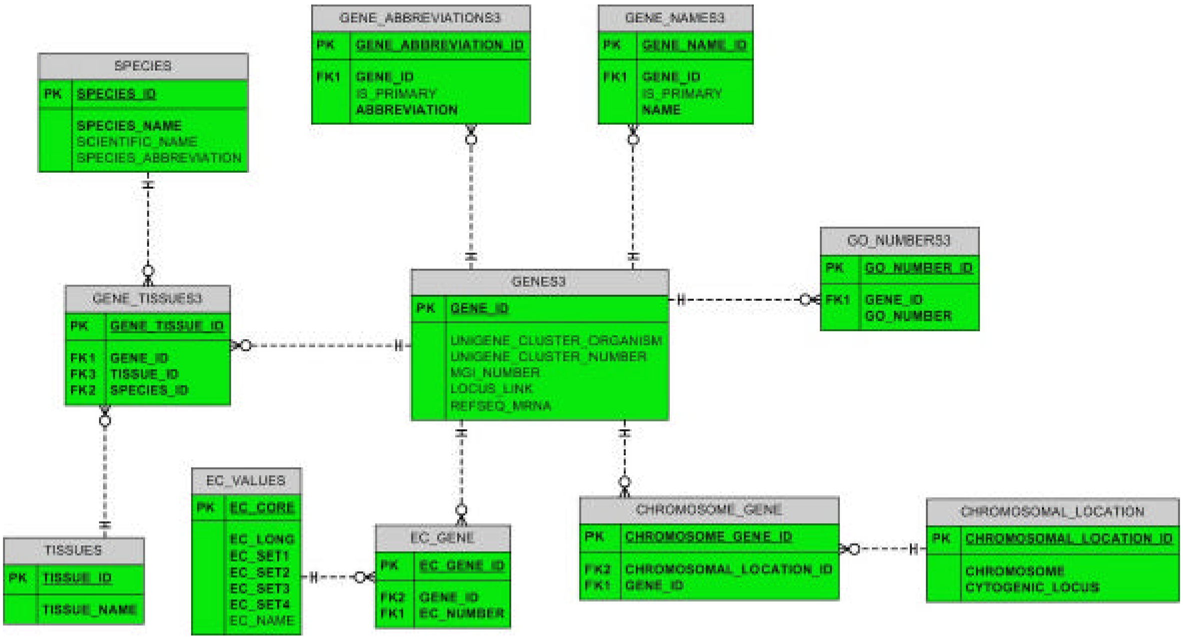


Figure 4. Relational database model for Gene Subsystem catalogue.
The Genes Subsystem catalogues all of the genes and indexes various annotative information. Annotative data include the following Gene ontology and Enzyme commission number information. PK indicates primary key unique to the table, FK indicate foreign key and dashed line indicates relationship between table s. Figure obtained from (http://dbzach.fst.msu.edu/dbZACH_info/dbZACH_FAQ.html).

In addition to being relatively easy to create and access, a relational database has the important advantage of being easy to extend. After the original database creation, a new data category can be added without requiring that all existing applications be modified. The standard user and application program interface to a relational database is the structured query language (SQL). SQL statements are used both for interactive queries for information from a relational database and for gathering data for reports. Some of the best-known RDBMS's include Oracle, Informix, Sybase, PostgreSQL and Microsoft Access.

3.1.1.1 Advantages of relational model

Advantages of relational data model include: (i) *Data independence* that provides a clear boundary between the logical and physical aspects of database management. (ii) *Simplicity*, providing a simpler structure. A simple structure is easy to communicate to users and programmers and a wide variety of users in an enterprise can interact with a simple model. (iii) *Extendability* In addition to being relatively easy to create and access, a relational database has the important advantage of being easy to extend.

After the original database creation, a new data category can be added without requiring that all existing applications be modified. (iv) The most common RDBMS like MySQL and PostgreSQL are freely available and can handle very large data.

Considering the above advantages, we built RETINOBASE using Relational Database Management System (PostgreSQL).

3.2 Biological databases

Several decades ago, scientists started to set up biological data collections for the centralized management of and easy access to experimental results, and to ensure long-term data storage and availability (Figure 5). Many early data collections were initially administered using word processing or spreadsheet applications. Database systems have been mainly developed and exploited in other scientific domains (such as physics, population or climatic studies, etc). Recently, the relatively new fields of functional genomics and system-level biology have increased the requirements of more evolved, elaborated and highly organized life-science databases. The most important tool for reaching an understanding of biology at the systems level is the analysis of biological models (Figure 5). The basic building blocks for these models are existing experimental data, which are stored in literally thousands of databases (Augen, 2001; Pennisi, 2005). As a consequence, database integration is a fundamental prerequisite for any study in systems biology (Ge *et al.*, 2003).

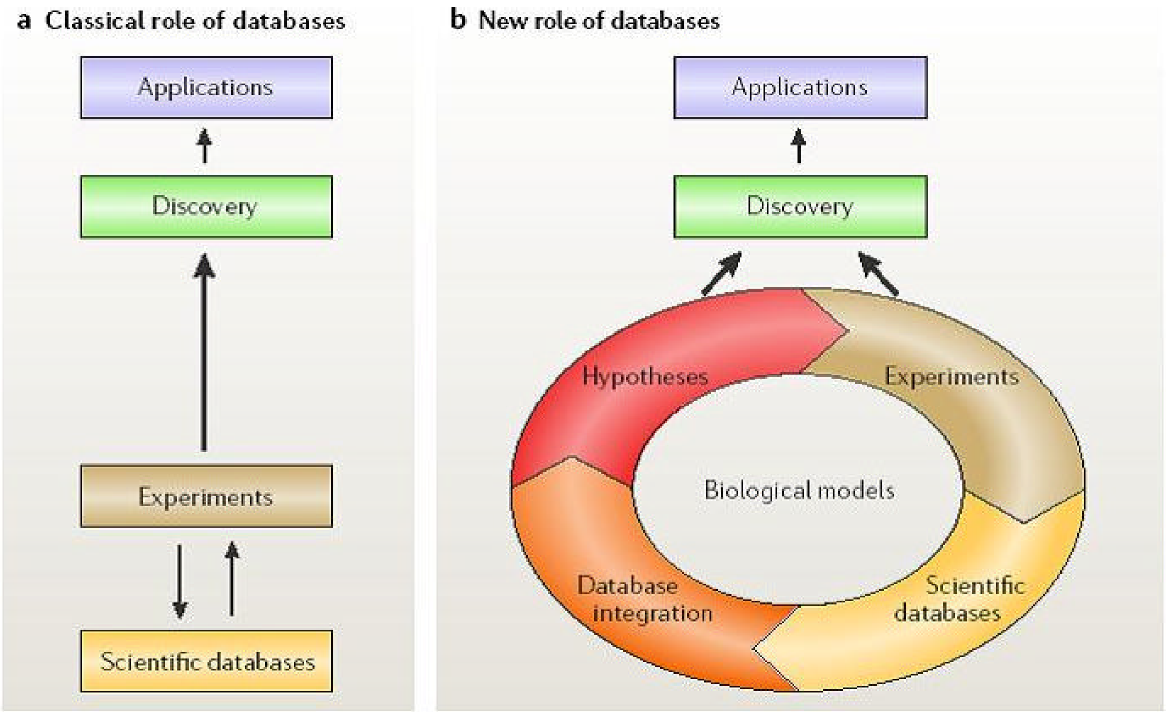


Figure 5. Classical and systems biology roles of life-science databases. The classical role of life-science databases is to provide easy access to and long-term storage of experimental results, with centralized data management. By contrast, more recent systems biological approaches exploit the information in databases to generate hypotheses for *in silico* discovery, which, after experimental verification, can be used to populate other databases. (Philippi and Kohler, 2006)

3.2.1 Specific features of biological databases

- Biological data is highly complex when compared to most other domains or applications. Definitions of such data must thus be able to represent a complex substructure of data as well as of relationships, to ensure that information is not lost during biological data modeling. Biological information systems must be able to represent any level of complexity in any data schema, relationship, or schema substructure.
- Biological data is large-scale and heterogeneous. Therefore, the systems handling biological data should be flexible in terms of data types and values. Constraints on data types and values should be used with care, since the unexpected values (e.g. outliers), which are common in biological data could be excluded, resulting in a loss of information.
- Database models (schemas) in biological databases change rapidly. This requires improved information flow between various database releases, as well as schema evolution and data object migration support. In most relational and

object database systems, the ability to extend the schema is not supported. What currently happens is that many biological/bioinformatics databases (such as GenBank, for example) release the entire database with new schemas once or twice a year rather than incrementally change the system.

- ❑ Representations of the same data by different biologists are likely to be different (even using the same system). Thus, it is necessary to have mechanisms for the alignment of different biological schemas.
- ❑ Most users of biological data need read-only access, and write access to the database is not required. Usually curators of the databases are the ones who need write access privileges. The vast majority of users generate a wide variety of read-access patterns into the database, but these patterns are not the same as those seen in traditional relational databases. User requested searches demand indexing of often-unexpected combinations of data classes.
- ❑ Most biologists do not have knowledge of the internal structure of the database or about its schema design. Biological database interfaces should display information to users in a manner that is applicable to the problem they are trying to address and that reflects the underlying data structure in an easily understandable manner.
- ❑ The context of data provides added meaning for its use in biological applications. Therefore it is important that the context is maintained and conveyed to the user when appropriate. It is also advantageous to integrate as many contexts as possible to maximize the interpretation of the biological data. For instance, a DNA sequence is not very useful without information describing its organization, function, etc.
- ❑ Defining and representing complex queries is extremely important to the biologist. Hence, biological systems must support complex queries and provide tools for building such queries.
- ❑ Users of biological information often require access to "old" values of the data, particularly when verifying previously reported results. Therefore, the changes in values must be supported through archiving to enable researchers to reconstruct previous work and re-evaluate prior and current information.

All these specific characteristics of biological data point to the fact that traditional DBMSs do not fully satisfy the requirements of complex biological data.

3.2.2 Database challenges

The integration of multiple heterogeneous sources of data as well as their management and dissemination to numerous users with diverse interests raises many challenges (Hernandez T, 2004).

- **Data collection and harmonisation**

Data related to protein and genomic sequences have been collected for many years and even though thousands of contributors and high throughput projects are now exponentially increasing the huge amount of available information, some common file formats and shared tools are widely used by the community, providing standards at least at the syntactic level. Nevertheless despite the continuing efforts to provide standardisation, indexation and public integration tools (Cheung *et al.*, 2005), the integration and cross querying of different sources is not yet simple, or even possible, particularly when using automatic processes.

The problems are compounded in the case of biomedical data because they were initially created for human experts and are generally extremely variable and highly unstructured.

- **Data integration**

Three major mechanisms of data storage can be envisaged: warehousing, indexation or federated databases (Nagarajan R, 2004). These mechanisms help in integrating many databases and facilitate rapid retrieval of specific data from large volumes of information. Warehousing involves the storage of all relevant information in a centralised database system. This is the simplest way to fully exploit the collected data because many processes can be run, which in turn allow complex data mining, standardisation and correlation studies. The main disadvantage of such an approach is that it takes lot of time and computational resources to load, transform and extract the data. In contrast, the indexation method (storing only links to the remote information) is very flexible and lightweight, but this complicates querying.

The federated database is a mixed approach that combines the advantages of these two methods. It consists of a local virtual relational database that maps the schemas of heterogeneous remote relational databases. The tables, columns, privileges, etc. are recognised by the federation system and cross queries joining the remote tables can be performed. The complete data network is then seen as a unique, local relational database. This necessitates, of course, a complex federation system able to integrate other database schemas and to deploy automatically the database queries. In addition, although most of the commonly used relational databases can be queried, specific wrappers need to be developed for other systems.

- **Dissemination, querying and visualization**

In setting up an efficient web access to the database, a certain number of aspects should be taken into consideration. A user friendly website should first, offer a simple and flexible querying system that allows the user to easily perform complex queries over the entire set of cross-linked data. Secondly, take into account the different interests of the users, their needs and access rights. An attractive design and an easy, intuitive navigation system and visualisation tools are crucial for an efficient exploitation of the databases and can only result from a close collaboration between all the database users.

Furthermore, if a large part of the available data needs to be exploited by external programs, the system must be computer-readable. Direct access through sockets, web services and API (Application Programming Interface) is necessary to allow an efficient use of the database. This automatic access also means that standards must be established at each level from data collection to dissemination.

3.3 Classification of biological databases

Currently, there are over 1,000 public and commercial biological databases. The most common biological databases contain sequences (nucleotide or amino acid), structures, gene or protein expression data, pathways and ontologies. Furthermore information about function, structure, chromosomal localisation, clinical effects of mutations as well as similarities of biological sequences can be found. Biological databases have become an important tool in assisting scientists to understand and explain a host of biological phenomena from the structure of biomolecules and their interactions, to the whole metabolism of organisms or the evolution of species. This knowledge contributes to the fight against diseases, assists in the development of medications and in discovering basic relationships amongst species in the history of life.

The biological databases can be roughly classified into the following groups:

General databases include nucleotide sequence databases such as GenBank (from the National Center for Biotechnology Information) (Benson *et al.*, 2007), protein sequence databases (such as the Uniprot database) (Apweiler *et al.*, 2004) or structure databases like PDB (Protein Data Bank) (Kouranov *et al.*, 2006), Gene Ontology databases like AmiGO.

Meta-databases include Entrez [National Centre for Biotechnology Information], euGenes (Indiana University), GeneCards (Weizmann Institute), SOURCE (Stanford University), Harvester (EMBL Heidelberg) and others. These federated databases can be considered as databases that harbour different databases. They collect information from different sources and usually make them available in a new and more convenient form. For example, the aim of Entrez, which is one of the most popular database querying systems, is to integrate most of the biological knowledge (such as scientific literature, DNA and protein sequence databases, 3D protein structure and protein domain data, population study datasets, expression data, assemblies of complete genomes, and taxonomic information) into a tightly interlinked system. It is mainly a retrieval system designed for searching its linked databases. Most of the databases are supported by both object and relational database management systems.

Specialized databases contain information concerning specific issues. Since the major public databases need to store data in a generalized fashion, often these databases do not support specialized requirements. To address this, many smaller, specialized databases have emerged. These databases, which contain information ranging from strain crosses to gene expression data, provide a valuable adjunct to the more visible public sequence databases, and the user is encouraged to make intelligent use of both types of databases in their searches. For example, resources like CGED (Kato *et al.*, 2005), SIEGE (Shah *et al.*, 2005) and GeneAtlas (Su *et al.*, 2002) are specialized databases that address specific problems related to gene expression; CGED addresses gene expression in various human cancer tissues, SIEGE focuses on epithelial gene expression changes induced by smoking in humans and GeneAtlas provides the expression profiles of genes in various mouse and human tissues.

3.4 Important databases for molecular biology

This section presents a brief description of the databases that have been used extensively in this thesis for data mining and extraction. Some of the databases have distinct database management system architecture, representing an important breakthrough in biology.

3.4.1 Nucleic acid sequence databases

The International Nucleotide Sequence Database Collaboration (INSDC, http://insdc.org) consists of a joint effort to collect and disseminate databases containing DNA and RNA sequences. It involves the following databases: DNA Data Bank of Japan (DDBJ, Japan) (Sugawara *et al.*, 2007), GenBank (USA) (Benson *et al.*, 2007) and the EMBL Nucleotide Sequence Database (European Molecular Biology Laboratory, Germany) (Kulikova *et al.*, 2007). New and updated data on nucleotide sequences contributed by research teams to each of the three databases are synchronized on a daily basis through continuous interaction between the collaborating organizations.

DDBJ began DNA data bank activities in earnest in 1986 at the National Institute of Genetics (NIG) with the endorsement of the Ministry of Education, Science, Sport and Culture. DDBJ is supported by RDBMS and is the sole DNA data bank in Japan, which is officially certified to collect DNA sequences from researchers and to issue the internationally recognized accession number to data submitters (Sugawara *et al.*, 2007).

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. The EMBL-Bank data is maintained in a relational database management system (RDBMS). The main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications (Kulikova *et al.*, 2007).

GenBank is the NIH (National Institute of Health) genetic sequence database, an annotated collection of all publicly available DNA sequences. The GenBank sequence entries are stored and maintained in the Sybase relational database management system (RDBMS). GenBank nucleotide records are available in the divisions CoreNucleotide, dbEST or dbGSS and can be searched in Entrez together or independently. The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information (Benson *et al.*, 2007).

3.4.2 Protein sequence databases: the UniProt databases

UniProt (Universal Protein Resource) is the world's most comprehensive catalogue of information on proteins. It is a central repository of protein sequence and function data created by joining the information contained in Swiss-Prot, TrEMBL, and PIR (Protein Information Resource)**.**

Organization of UniProt Databases

The UniProt databases consist of three database layers each optimized for different uses:

- ❑ The **UniProt Archive (UniParc)** provides a stable, comprehensive sequence collection without redundant sequences by storing the complete body of publicly available protein sequence data.
- ❑ The **UniProt Knowledgebase (UniProtKB)** (Apweiler *et al.*, 2004) is the central database of protein sequences with accurate, consistent, and rich sequence and functional annotation.
- ❑ The **UniProt Reference Clusters (UniRef)** databases provide non-redundant reference clusters based on the UniProt knowledgebase (and selected UniParc records) in order to obtain complete coverage of sequence space at several resolutions.

UniProtKB (UniProt KnowledgeBase) which stands on relational database management system (RDBMS) is the central hub for the collection of functional information on proteins with accurate, consistent and rich annotation. In addition to obtaining the core data mandatory for each UniProtKB entry (principally, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added. This includes widely accepted biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data. Created by merging the data in Swiss-Prot, TrEMBL and PIR-PSD, individual UniProt KnowledgeBase entries may contain more information than was available in any given separate source database. The UniProtKB consists of two sections:

- UniProtKB/Swiss-Prot; a curated protein sequence database which strives to provide a high level of validated annotation (such as the description of the function of a protein, its domain structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.
- UniProtKB/TrEMBL; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

3.4.3 Protein structure database

Protein Data Bank (PDB)

The Protein Data Bank (PDB) (Kouranov *et al.*, 2006) is the single worldwide depository of information about the three-dimensional structures of large biological molecules, including proteins and nucleic acids. A variety of information associated with each structure is available through the Research Collaboratory for Structural Bioinformatics (RCSB) PDB including sequence details, atomic coordinates, crystallization conditions, 3-D structure neighbours computed using various methods, derived geometric data, structure factors, 3-D images and a variety of links to other resources. At present, PDB has about 47137 structures (November, 2007) (Figure 6).

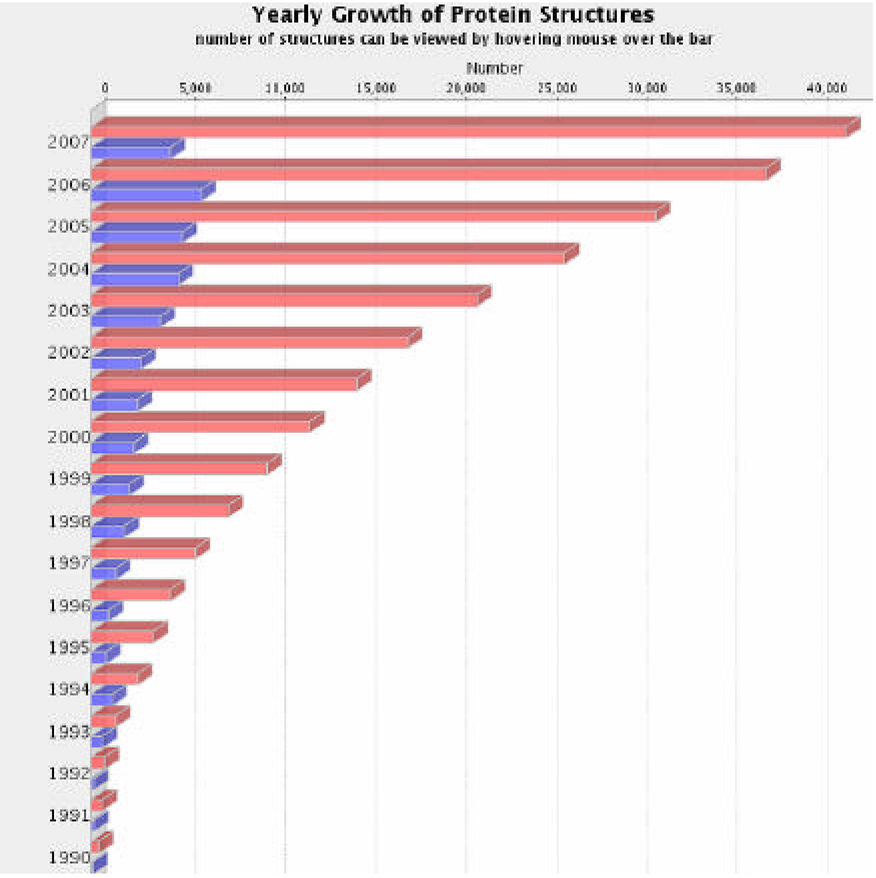


Figure 6. Yearly growth of protein structures in PDB.

The PDB database architecture is built on 5 major components: (Berman *et al.*, 2000)

- The core PDB **relational database** contains all deposited information in a tabular form that can be accessed across any number of structures.
- The final curated data files (in PDB and mmCIF formats) and data dictionaries are the archival data and are present as ASCII files in the ftp archive.
- The POM (**P**roperty **O**bject **M**odel)-based databases, which consist of indexed objects containing native (e.g., atomic coordinates) and derived properties (e.g., calculated secondary structure assignments and property profiles).
- The Biological Macromolecule Crystallization Database (BMCD; is organized as a **relational database** within the database management system Sybase and contains three general categories of literature derived information: macromolecular, crystal and summary data.

- The Netscape LDAP server is used to index the textual content of the PDB in a structured format and provides support for keyword searches.

3.4.4 Pathway database: KEGG database

KEGG (Kyoto Encyclopaedia of Genes and Genomics) (Kanehisa *et al.*, 2006) (http://www.genome.jp/kegg/) is a database of biological systems, consisting of genetic building blocks of genes and proteins (KEGG GENES), chemical building blocks of both endogenous and exogenous substances (KEGG LIGAND), molecular wiring diagrams of interaction and reaction networks (KEGG PATHWAY), and hierarchies and relationships of various biological objects (KEGG BRITE). KEGG provides a reference knowledge base for linking genomes to biological systems and also to environments by the processes of PATHWAY mapping and BRITE mapping.

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for:

- Metabolism: carbohydrate, energy, lipid, nucleotide, amino acid, glycan, cofactor/vitamin, secondary metabolite and xenobiotics
- Genetic information processing
- Environmental information processing
- Cellular processes
- Human diseases
- Drug development

At present KEGG PATHWAY has 56,987 pathways generated from 333 reference pathways. Representation of retinol metabolism pathways in animals in KEGG PATHWAY is presented in (Figure 7).

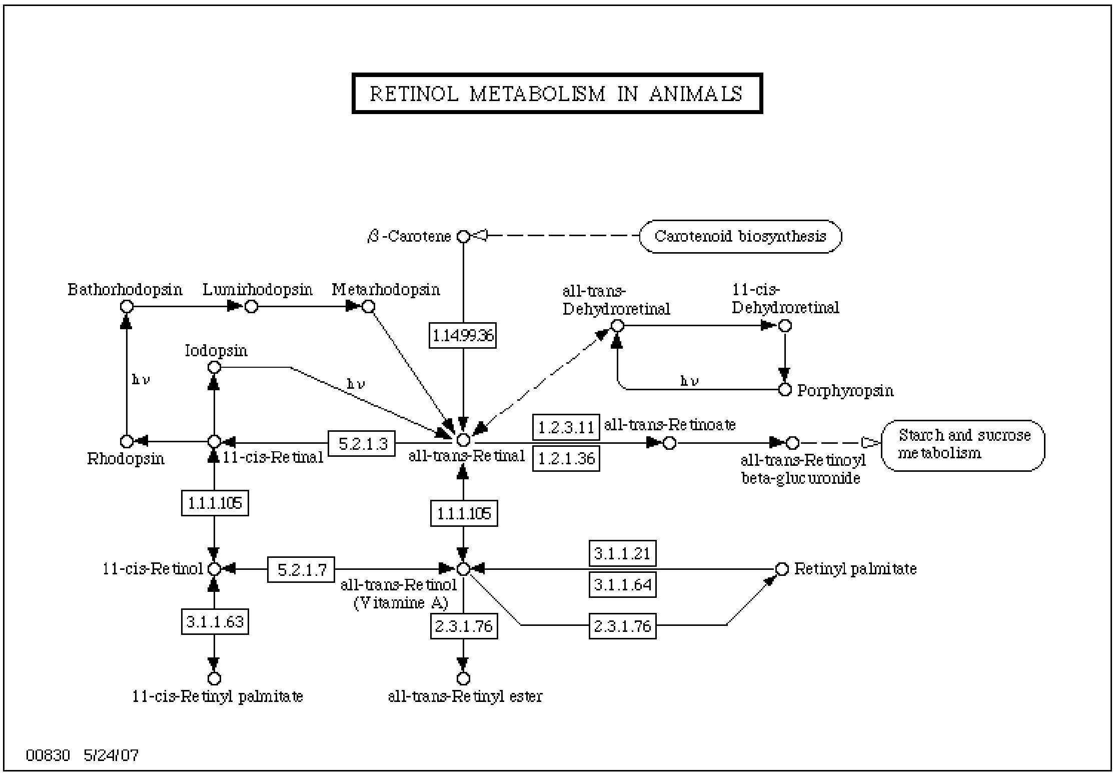


Figure 7. Representation of retinol metabolism in animals in KEGG PATHWAY database.

Indicating different interactions and relationship between the various molecules in the pathway. Number indicates enzyme commission number; line and broken line indicate molecular interaction and indirect affect respectively, oval boxes indicate another KEGG maps and 'h?' represents photon activation.

3.4.5 Gene Ontology database

Biological knowledge is inherently complex and so cannot readily be integrated into existing databases of molecular (for example, sequence) data. Ontology is the concrete form of a conceptualisation of a community's knowledge of a domain. Ontology is a formal way of representing this knowledge, in which concepts are described both by their meaning and their relationship to each other. One way of capturing knowledge within bioinformatics applications and databases is the use of such ontologies.

The ontologies grouped together at the OBO web site cover a wide range of biomedical fields, such as specific organism anatomies, taxonomic classifications or transcriptomic and proteomic experimental protocols and data. Various ontologies have also been developed for particular aspects of single sequences, such as gene structure (SO) (Eilbeck *et al.*, 2005) protein function (GO) or protein–protein interactions (MI) (Hermjakob *et al.*, 2004). Some work has also begun to develop standard data formats to represent RNA sequences and structures and the RNA Ontology Consortium (ROC) (Leontis *et al.*, 2006)) has been established to build a formal ontology. The OBO ontologies can be accessed from a single location with a unified output format, using the Ontology Lookup Service (OLS) at http://www.ebi.ac.uk/ontology-lookup/.

The most widely known and used ontology is the Gene Ontology (GO). The GO project members (http://www.geneontology.org) (Ashburner *et al.*, 2000) have developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated **biological processes, cellular components and molecular functions** in a species-independent manner (Figure 8).

The GO Database is a relational database housing both the Gene Ontology and the annotations of genes and gene products to terms in the GO. The advantage of housing both ontologies and annotations in a single database is that powerful queries can be performed over annotations using the ontology.

The GO Database forms the base of the AmiGO (is the official tool for searching and browsing the Gene Ontology database) browser and search engine. It is built from source data at regular intervals, and is currently housed as a MySQL database. There are three separate aspects to this effort: first, the development and maintenance of the ontologies themselves; second, the annotation of gene products, which entails making associations between the ontologies and the genes and gene products in the collaborating databases; third, the development of tools that facilitate the creation, maintenance and use of ontologies.

The use of GO terms by collaborating databases facilitates uniform queries across them. The controlled vocabularies are structured so that they can be queried at different levels. This structure also allows annotators to assign properties to genes or gene products at different levels, depending on the depth of knowledge about that entity.

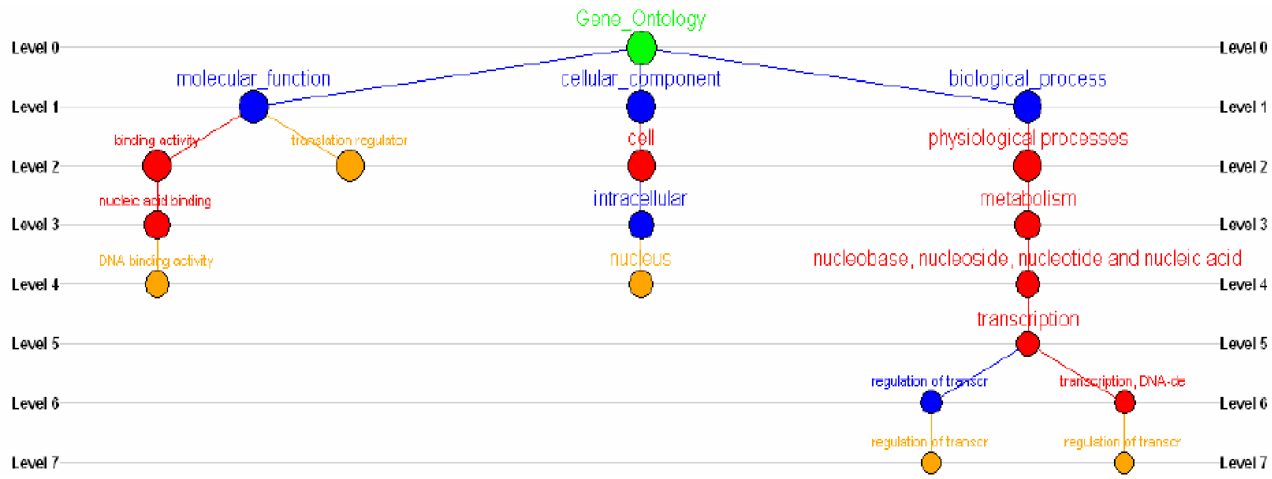


Figure 8. Example of gene annotation with GO.
General representation indicating different levels of classification, three structured and controlled vocabularies, Molecular function, cellular component and biological process for transcriptional regulator.

3.4.5.1 DAVID bioinformatic resources

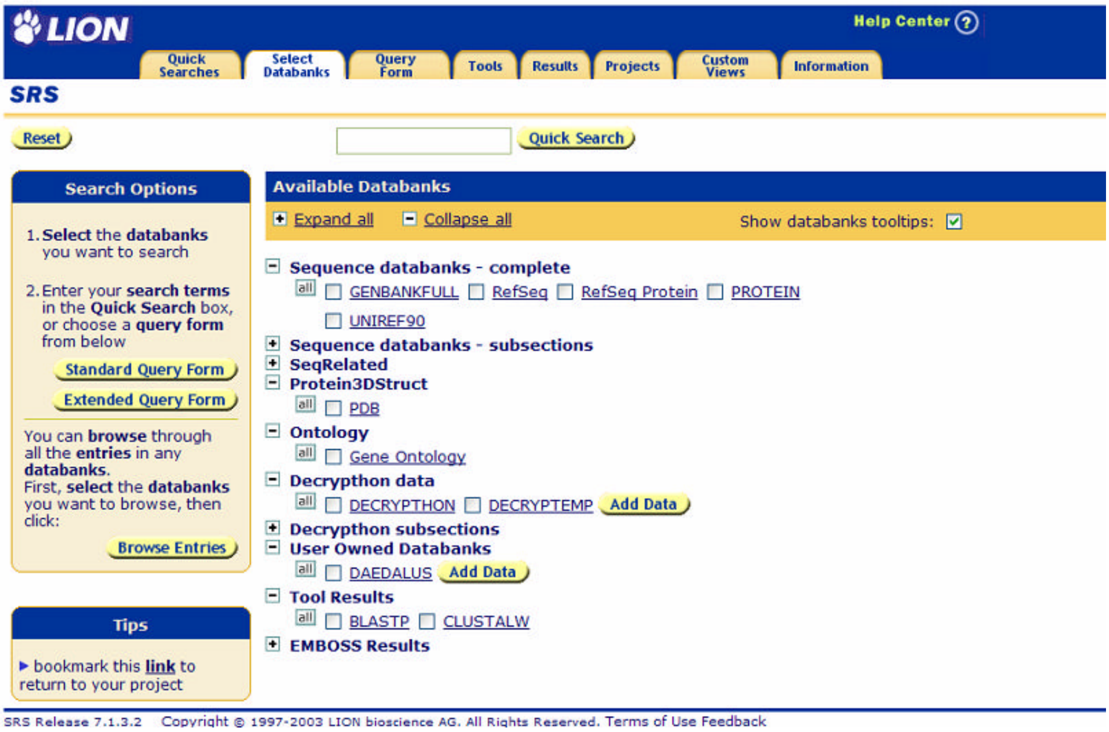
The Database for Annotation, Visualization and Integrated Discovery (DAVID) is a web-based functional annotation tool, particularly for gene-enrichment analysis on DAVID knowledgebase. This is designed on the "DAVID Gene Concept", a graph theory evidence–based method to agglomerate heterogeneous and widely distributed public databases. It also provides an enhanced set of bioinformatics tools, not only limited to functional annotations, to systematically summarize relevant biological patterns from user-classified gene list. This helps users to quickly understand the biological themes under the study. Tools provided by DAVID include Functional annotation tool (perform gene-enrichment analysis, pathway mapping, gene/term similarity search, graphic representation, homologue match), Gene Accession Conversion Tool (converts a list of gene IDs, accessions to others of users choice). The DAVID Gene Concept method groups tens of million of identifiers from over 65,000 species into 1.5 million unique protein and gene records. The grouping of such identifiers allows agglomeration of a divers array of functional and sequence annotation, greatly enriching the level of biological information available for a given gene (Dennis *et al.*, 2003).

Gene ontology annotations in DAVID are divided into 5 levels. Level 1 is a general description whereas level 5 is a more specific description. A term at level 5 is a child of term at level 1 for a given gene. Level 1 provides the highest list coverage with the least amount of term specificity. With each increasing level coverage decreases while the specificity increases so that level 5 provides the least amount of coverage with the highest term specificity.

3.5. Sequence Retrieval Software (SRS)

SRS is a powerful unified interface to different biological databases including nucleic acid and protein sequences, structures, protein domains and metabolic pathways. It has been developed into an integration system for both data retrieval and sequence analysis applications. It is designed for the extraction of semi-structured data, i.e. textual data with a pre-defined structure that may include redundancies or irregularities.

The relational module provides a choice of maintaining data sources in either a relational format, or as flat-files or as a combination of the two. This is facilitated by SRS's use of automatic schema modeling, whereby the database schema is analyzed, and the relational queries are created dynamically from the SRS query, without the need for manual programming. SRS performs a grammatical parsing of the information contained in the flat files and then indexes the different fields associated with each entry. This indexing allows a rapid access to the entry fields via complex queries, as well as cross-referenced queries that exploit the links between the different databases. Version 7.1.3.2 of SRS is currently installed at the laboratory. Database queries can be performed using UNIX commands or interactively (http://bips.u-strasbg.fr/srs/, (Figure 9).



1. Select the databanks you want to search

2. Enter your search terms in the Quick Search box, or choose a query form from below

You can browse through all the entries in any databanks. First, select the databanks you want to browse, then click:

▸ bookmark this link to return to your project

SRS Release 7.1.3.2 Copyright © 1997-2003 LION bioscience AG. All Rights Reserved. Terms of Use Feedback

Figure 9. Screenshot of SRS web server at IGBMC.

3.5.1 Entrez-National Center for Biotechnology Information (NCBI)

Entrez (Geer and Sayers, 2003) is the integrated, text-based search and data retrieval system developed by the National Center for Biotechnology Information (NCBI) (Wheeler *et al.*, 2007) that provides integrated access to a wide range of data domains, (literature, nucleotide and protein sequences, complete genomes, three-dimensional structures… that are linked in a relational schema). Entrez includes powerful search features that retrieve not only the exact search results but also related records within a data domain that might not be retrieved otherwise and associated records across data domains. Entrez provides a single access point to previously disparate data.

An Entrez “node” (Figure 10) is a collection of data that is grouped together and indexed together. It is usually referred to as an Entrez database. In the first version of Entrez, there were three nodes: published articles, nucleotide sequences and protein sequences. Each node represents specific data objects of the same type, e.g., protein sequences, which are each given a unique ID (UID) within that logical Entrez Proteins node. Records in a node may come from a single source (e.g., all published articles are from PubMed) or many sources (e.g., proteins are from translated GenBank sequences, SWISS-PROT, or PIR).

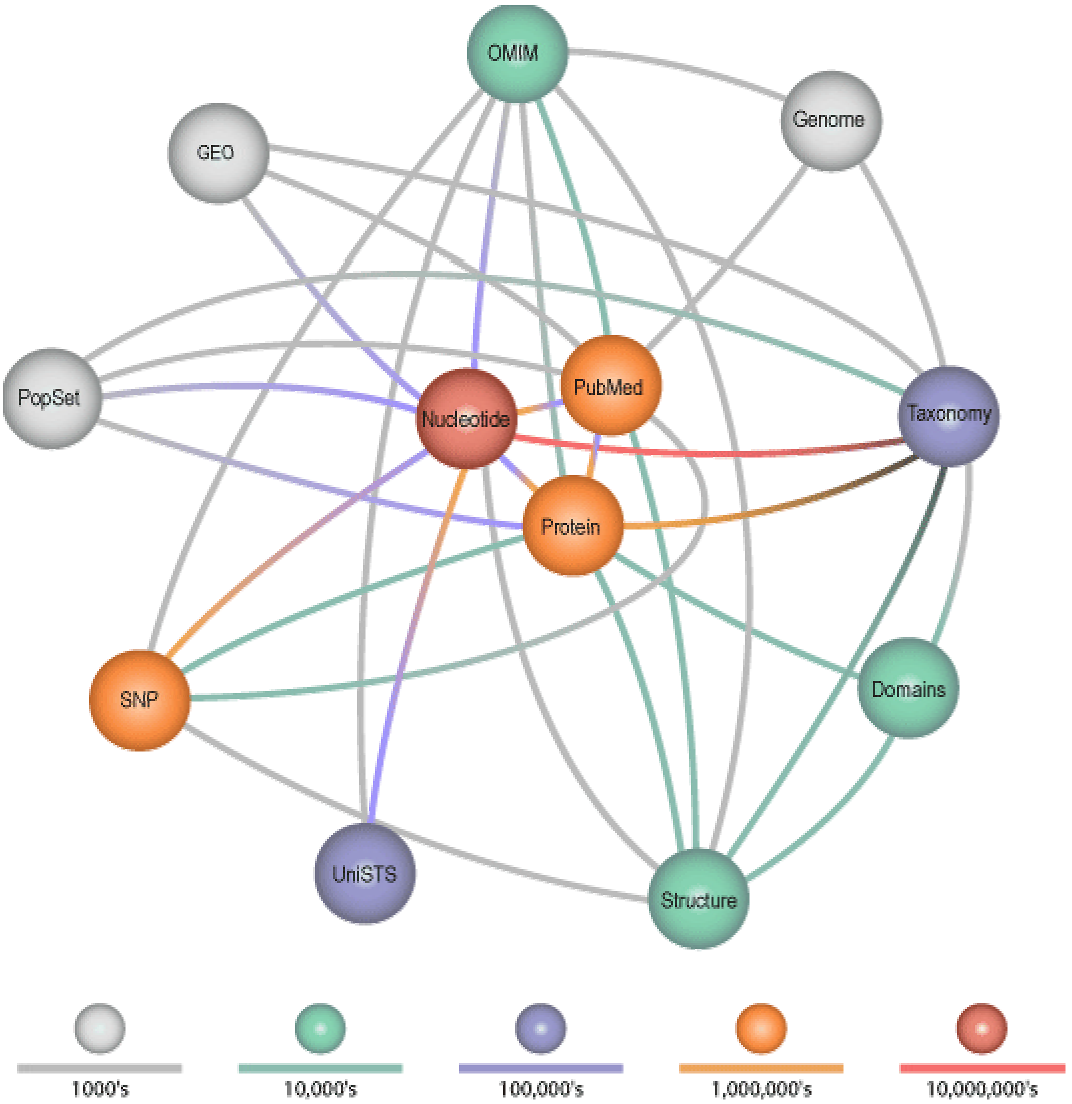


Figure 10. Entrez nodes
The original version of Entrez had just 3 nodes: nucleotides, proteins, and PubMed abstracts. Entrez has now grown to nearly 20 nodes. This kind of node association system is especially efficient in handling and retrieving large, rapidly changing biological data sets. A PopSet is a set of DNA sequence that has been collected to analyse the evolutionary relatedness of the population. GEO (Gene Expression Omnibus), OMIM (Online Mendelian Inheritance in Man), Line indicates direct link between 2 nodes and if there is no line the 2 nodes are connected through other nodes. Figure adapted from http://www.ncbi.nlm.nih.gov/Database/datamodel/index.html.

3.6 UCSC genome browser for gene localization

The University of California Santa Cruz (UCSC) Genome Browser Database is an up to date source for genome sequence data integrated with a large collection of related annotations. The database is optimized to support fast interactive performance with the web-based UCSC Genome Browser, a tool built on top of the database for rapid visualization and querying of the data at many levels. The annotations for a given genome are displayed in the browser as a series of tracks aligned with the genomic sequence. Sequence data and annotations may also be viewed in a text-based tabular format or downloaded as tab-delimited flat files.

Sequence and annotation data for each genome assembly are stored in a **MySQL relational database**, which is quite efficient at retrieving data from indexed files. The database is loaded in large batches and is used primarily as a read-only database. To improve performance, each of the Genome Browser web servers has a copy of the database on its local disk.

With such a large data set, care must be taken in the database organization to ensure good interactive performance in the Genome Browser, which creates each line of its graphical annotation image by querying the corresponding table in the database. The database is optimized to support the browser's range-based queries, with additional optimizations to accommodate the varying sizes of the positional tables in the database (Kent *et al.*, 2002).

The UCSC genome browser interface with the human rhodopsin gene is presented in (Figure 11)

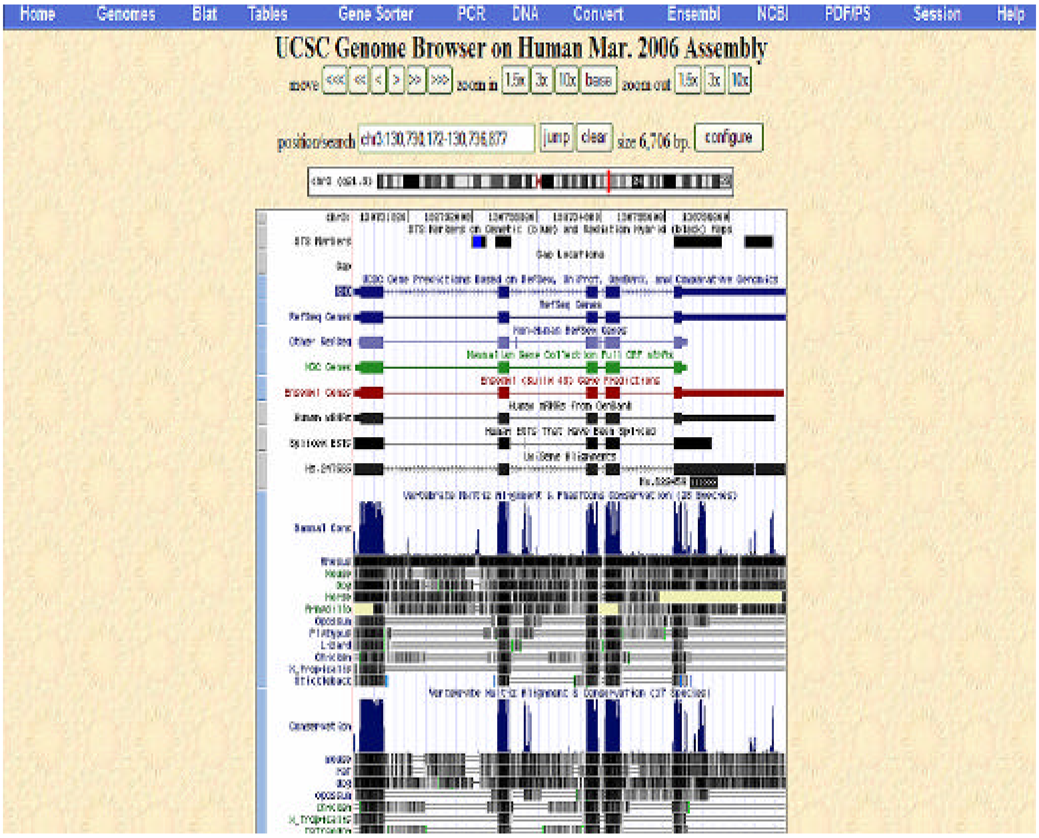


Figure 11. UCSC genome browser interface.
The human rhodopsin gene is shown in the UCSC Genome Browser. The mRNA, EST, the repeat localization as well as the conservation profile during evolution (from the MULTIZ alignment between 17 species) are depicted. http://genome.ucsc.edu/.

3.7 Microarray databases

3.7.1 Types of Microarray databases

- **General repositories**
- **Microarray database system**

3.7.1.1 General repositories

Databases such as Gene Expression Omnibus (GEO) and Array Express acts as general repositories to store gene expression data from a variety of sources such as tissues, cells and with variety of treatments performed on different arraytypes.

3.7.1.1.1 Gene Expression Omnibus (GEO)

The examination of gene expression using high-throughput methodologies has become very popular in recent years. Techniques such as microarray hybridization and Serial Analysis of Gene Expression (SAGE) allow the simultaneous quantification of tens of thousands of gene transcripts. The Gene Expression Omnibus (GEO) (Barrett *et al.*, 2007) is a public repository provided by the NCBI that archives and freely distributes high-throughput gene expression data submitted by the scientific community. Data types currently stored include, but are not limited to, cDNA and oligonucleotide microarrays that examine gene expression, SAGE, massively parallel signature sequencing (MPSS), Comparative Genomic Hybridization (CGH) array, and chromatin-immunoprecipitation on array (ChIP-chip) studies. GEO currently (November, 2007) stores approximately a billion individual gene expression measurements, derived from 6832 experiments performed over 3957 platforms and in over 100 organisms, addressing a wide range of biological issues. These huge volumes of data can be effectively explored, queried, and visualized using user-friendly Web-based tools. GEO is accessible at www.ncbi.nlm.nih.gov/geo.

The GEO database architecture (Barrett and Edgar, 2006) is designed for the efficient capture, storage, and retrieval of heterogeneous sets of high-throughput molecular abundance data. The structure is sufficiently flexible to accommodate evolving state of the art technologies. GEO have there own analysis tools which includes gene expression profile charts and “DataSet” clusters mainly related to clustering.

Data supplied by submitters are stored as three main entities in a MSSQL Server relational database:

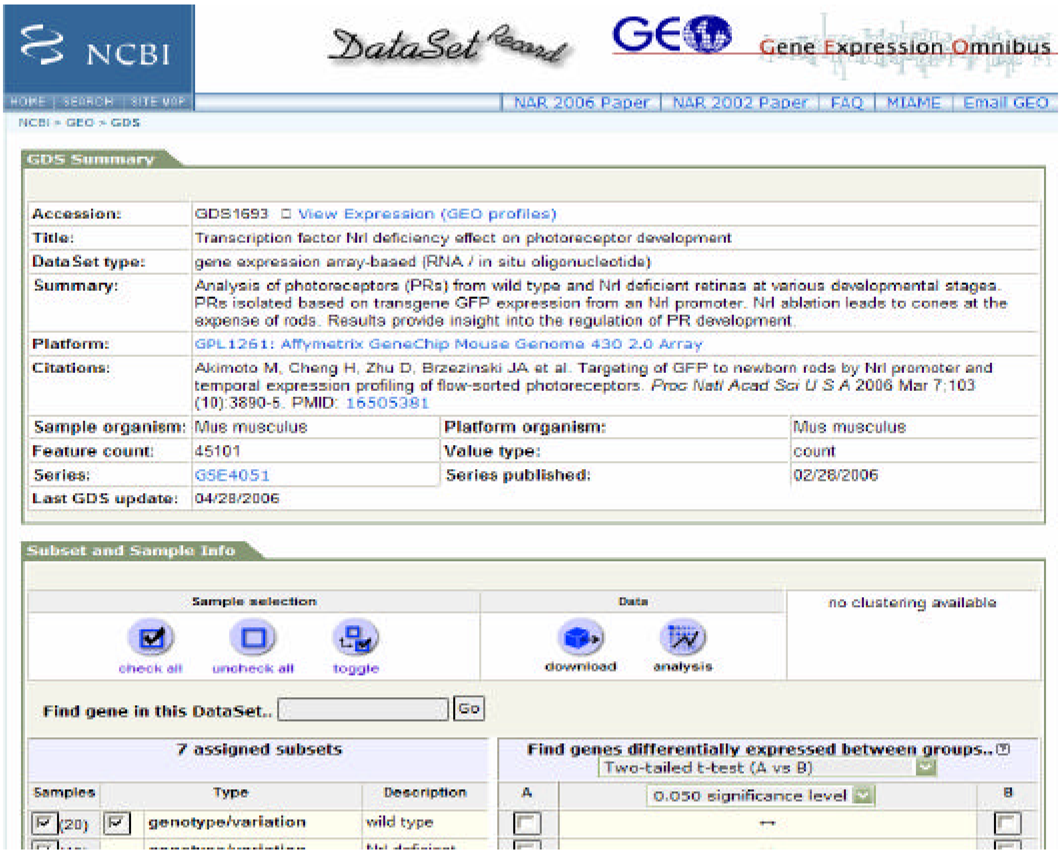
Platform: Includes a summary description of the array and a data table defining the array template. Each row in the table corresponds to a single element and includes sequence annotation and tracking information as provided by the submitter.

Sample: Includes a description of the biological source and the experimental protocols to which it was subjected, and a data table containing hybridization measurements for each element on the corresponding platform.

Series: Defines a set of related samples considered to be part of a study, and describes the overall study aim and design.

A DataSet in GEO provides two separate perspectives of the data:

1. An experiment-centered rendering that encapsulates the whole study. This information is presented as a "DataSet record". DataSet records comprise a synopsis of the experiment, a breakdown of the experimental variables, access to auxiliary objects, several data display and analysis tools, and download options (Figure 12).



Accession:	GDS1693 □ View Expression (GEO profiles)		
Title:	Transcription factor Nrl deficiency effect on photoreceptor development		
DataSet type:	gene expression array-based (RNA / in situ oligonucleotide)		
Summary:	Analysis of photoreceptors (PRs) from wild type and Nrl deficient retinas at various developmental stages. PRs isolated based on transgene GFP expression from an Nrl promoter. Nrl ablation leads to cones at the expense of rods. Results provide insight into the regulation of PR development.		
Platform:	GPL1261: Affymetrix GeneChip Mouse Genome 430 2.0 Array		
Citations:	Akimoto M, Cheng H, Zhu D, Brzezinski JA et al. Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors. *Proc Natl Acad Sci U S A* 2006 Mar 7;103 (10):3890-5. PMID: 16505381		
Sample organism:	Mus musculus	Platform organism:	Mus musculus
Feature count:	45101	Value type:	count
Series:	GSE4051	Series published:	02/28/2006
Last GDS update:	04/28/2006		

Samples		Type	Description	A		B
☑ (20)	☑	genotype/variation	wild type	□	↔	□

Figure 12. Screenshot of a typical DataSet record.
The record includes a summary of the experiment, links to related records and publications, subset designations and classifications, download options, and access to mining features such

as cluster heat maps and 'Query group A vs B' tool. Figure adapted from www.ncbi.nlm.nih.gov/geo.

2. A gene-centered rendering that presents quantitative gene expression measurements for one gene across a DataSet. This information is presented as a "GEO Profile". A GEO Profile comprises gene identity annotation, the DataSet title, links to auxiliary information, and a chart depicting the expression level of that gene across each Sample in the DataSet.

3.7.1.1.2 ArrayExpress

ArrayExpress is a public repository for microarray data hosted by European Bioinformatics Institue (EBI) (Parkinson *et al.,* 2007). The ArrayExpress data warehouse stores gene-indexed expression profiles from a curated subset of experiments in the repository, it also provides an environment to update the annotations for array and runs on Oracle RDBMS (Figure 13).

ArrayExpress has two major components:

- ❑ ArrayExpress experiment repository - the main database containing complete data supporting publications.
- ❑ ArrayExpress gene expression profile data warehouse - contains gene-indexed expression profiles from a curated subset of experiments from the repository.

Public data in ArrayExpress are made available for browsing and querying on experiment properties, submitter, species, etc. Queries return summaries of experiments and complete data, or subsets can be retrieved. A subset of the public data are re-annotated to update the array design annotation and curated for consistency. These data are stored in the data warehouse and can be queried on gene, sample, and experiment attributes. Results return graphed gene expression profiles, one graph per experiment. ArrayExpress is accessible at (http://www.ebi.ac.uk/Databases/microarray.html). Currently (November, 2007) ArrayExpress has around 4446760 profiles from 5597 hybridizations and in 180 experiments.

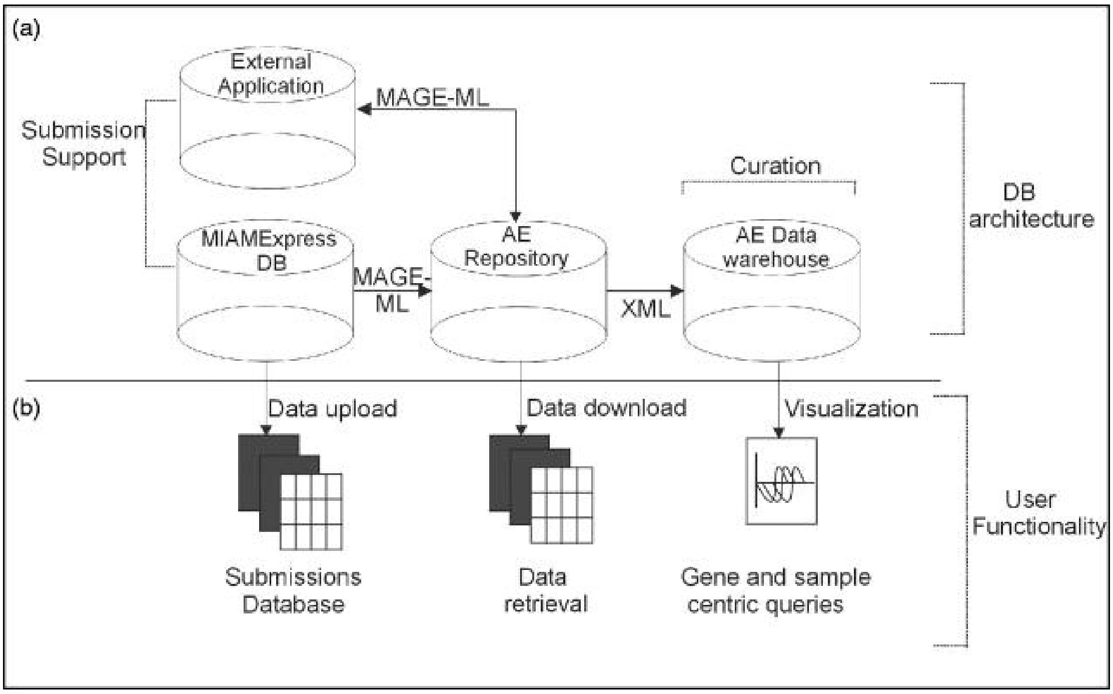


Figure 13. Architecture of ArrayExpress database.
a) Indicates ArrayExpress architecture and database activities b) Functionality experienced by the user is shown. (Parkinson *et al.*, 2005)

In general, both GEO (Barrett *et al.*, 2006; Barrett *et al.*, 2007) and ArrayExpress (Parkinson et al., 2007) are general repositories and operate as central data distribution centers and usually do not provide direct tools for integrated analysis across different experiments. Both databases are MIAME compliant (see below) and support both experiment and gene centric queries to visualize gene expression profiles.

3.7.1.2 Microarray database system

This represents a customizable solution for the management and analysis for all areas of microarray experimentation.

3.7.1.2.1 BioArray Software Environment (BASE)

BioArray Software Environment (BASE) (Saal *et al.*, 2002) provides an integrated framework for storing and analyzing microarray information. The software was developed on the GNU/Linux operating system (OS) in the PHP language, with data being stored in a relational database MySQL and communicated to the user through the Apache Webserver. The user interface employs Java and JavaScript in addition to plain HTML. The system integrates biomaterial information, raw images and extracted data, also provides a plug-in architecture for data transformation, data viewing and analysis modules. Data can be exported in a multitude of formats for local analysis and publication (Figure 14).

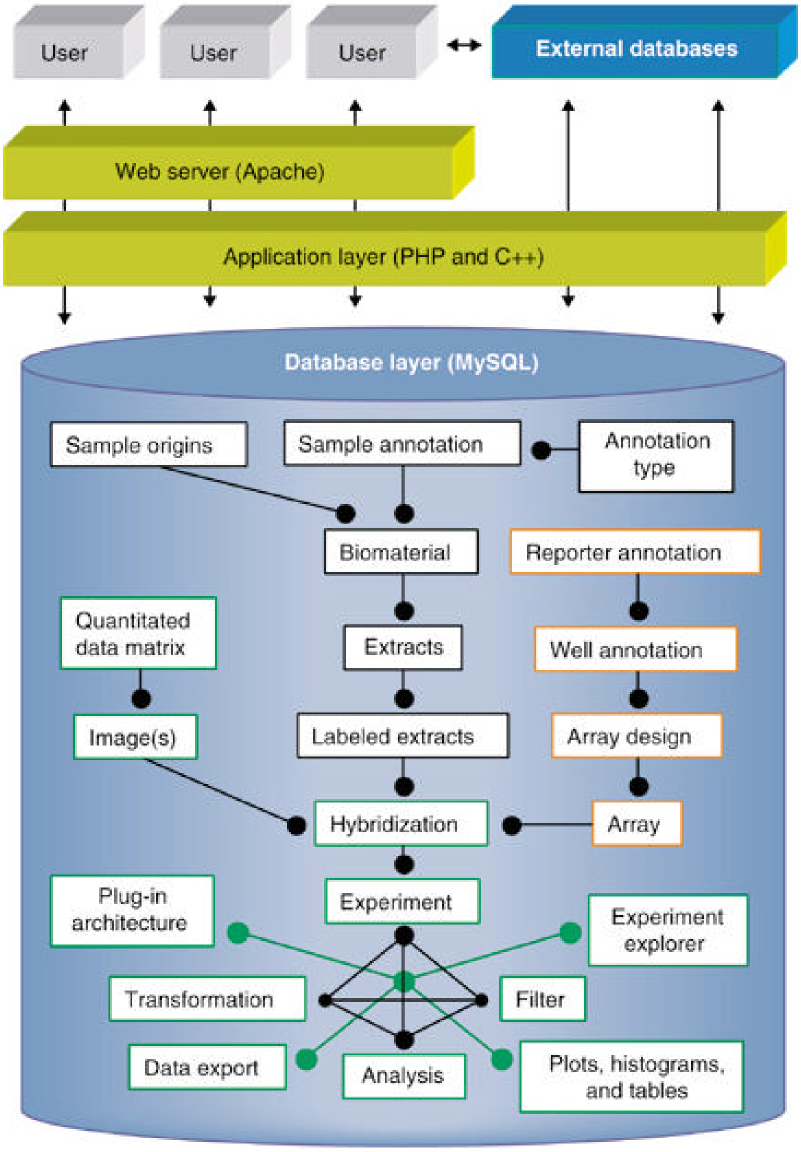


Figure 14: Simplified schematic overview of software structure of BASE. Arrows represent the flow of information. Closed circle connectors represent logical relationships between database classes. Database classes outlined by black boxes relate to biomaterials; array production LIMS items are highlighted by orange boxes; and data-analysis features are within green boxes.

3.7.1.2.2 Mediante

The Mediante (Le Brigand and Barbry, 2007) is a MIAME-compliant microarray data manager is a J2EE platform deployed under a Tomcat web server. It is based on a PostgreSQL relational database organized into four distinct related modules: (i) annotation of the transcripts, (ii) annotation of the probes, (iii) microarray production and (iv) information about biological experiments and hybridizations. Data can be easily exported from Mediante to other applications like statistical analysis software such as the Bioconductor environment of R and public repositories such as GEO and ArrayExpress.

Initially, RETINOBASE was planned to use database schema of BASE, due to unforeseen problems with data visualization module and further to support our own querying and visualization systems; we choose not to use database schema of base rather we developed our own relational schema for the RETINOBASE. However we used module related to sample annotation of BASE in our schema because this module was well defined and also suitable for our database.

Chapter 4. Transcriptome

This chapter describes the high throughput technologies used to analyse the transcriptome, i.e. the combined set of expressed RNA transcripts present in a cell at a given time. Nevertheless, before describing the transcriptomic approaches, the various levels of gene expression regulation, notably the transcriptional level, will be quickly overviewed to illustrate that RNA transcript amount is a composite resulting from numerous complex processes involved in the precise control of RNA transcripts synthesis and maintenance.

4.1 Regulation of gene expression

Regulation of gene expression occupies a central role in the control of the flow of genetic information from genes to proteins. Regulatory events on multiple levels ensure that the majority of the genes are expressed under controlled circumstances to yield temporally controlled, cell and tissue-specific expression patterns. As the differences among the various cell types of an organism depend on the particular genes that the cells expressed and/or are expressing, at what level is the control of gene expression exercised? There are many steps in the pathway leading from DNA to protein, and all of them can in principle be regulated. Schematically, a cell can control the gene products at various levels (Figure 15):

1. Controlling when and how often a given gene is transcribed (**Transcriptional control**).
2. Controlling how the RNA transcript is spliced or otherwise processed (**RNA processing control**).
3. Selecting which processed mRNAs in the cell nucleus are exported to the cytosol and determining where in the cytosol they are localized (**RNA transport and localization control**).
4. Selecting which mRNAs in the cytoplasm are translated by ribosomes (**Translational control**).
5. Selectively destabilizing certain mRNA molecules in the cytoplasm (**mRNA degradation control**).
6. Selectively activating, inactivating, degrading, or compartmentalizing specific protein molecules after they have been made (**Protein activity control**).

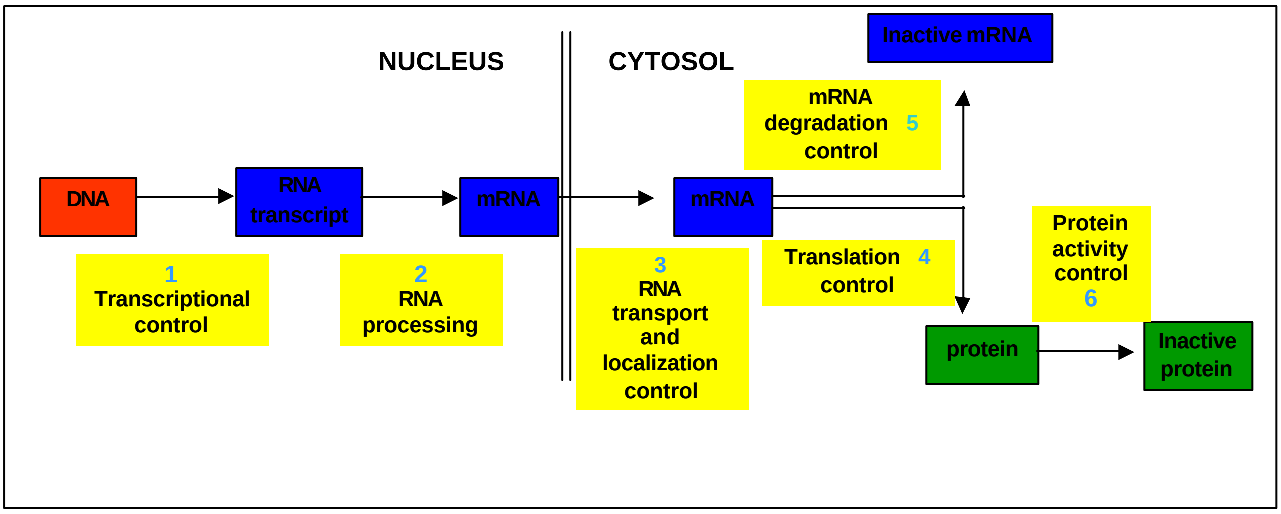

NUCLEUS
CYTOSOL
Inactive mRNA
mRNA degradation control 5
DNA
RNA transcript
mRNA
mRNA
1 Transcriptional control
2 RNA processing
3 RNA transport and localization control
Translation control 4
Protein activity control 6
protein
Inactive protein

Figure 15. Steps at which eukaryotic gene expression can be controlled.
Figure adapted from (Alberts, 2002).

4.1.1 Transcriptional control in eukaryotic cells

The packaging of eukaryotic DNA (Figure 16) into chromatin also provides opportunities for regulation. Indeed, in addition to their direct actions in assembling the RNA polymerase and the general transcription factors on DNA, gene activator proteins also promote transcription initiation by changing the chromatin structure of the regulatory sequences and promoters of genes. The two most important ways of locally altering chromatin structure are through covalent histone modifications and nucleosome remodelling. Many gene activator proteins make use of both mechanisms by binding to and thereby recruiting histone acetyl transferases (HATs), commonly known as histone acetylases, and ATP-dependent chromatin remodelling complexes to work on nearby chromatin. The local alterations in chromatin structure that ensue allow greater accessibility to the underlying DNA. This accessibility facilitates the assembly of the general transcription factors and the RNA polymerase holoenzyme at the promoter and allows the binding of additional gene regulatory proteins to the control region of the gene (Struhl, 1998).

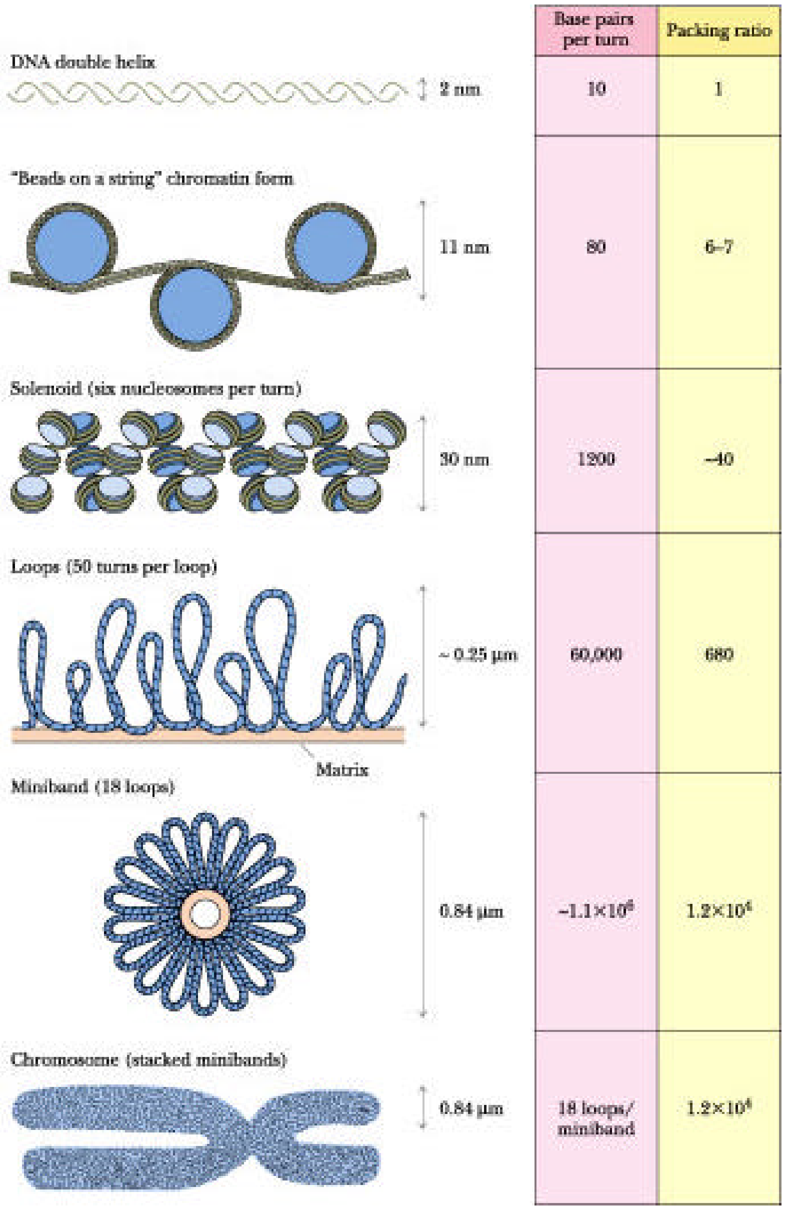


Figure 16. Different hierarchies in DNA packaging.
The DNA starts out as single strand double helices and is condensed until it reaches chromosomal structures.
Figure adapted from (http://www.web.virginia.edu/Heidi/chapter12/Images/8883n12_31.jpg).

The regulation of transcription is central and can be envisaged at two major levels. First, eukaryotes make use of gene regulatory proteins that can act even when they are bound to DNA thousands of nucleotide pairs away from the promoter that they influence. Second, RNA polymerase II, which transcribes all protein-coding genes, cannot initiate transcription on its own. It requires a set of proteins called general transcription factors, which must be assembled at the promoter before transcription can begin. This assembly process is a rate limiting step for transcription initiation, and many gene regulatory proteins influence these step (Kornberg, 1999) (Figure 17).

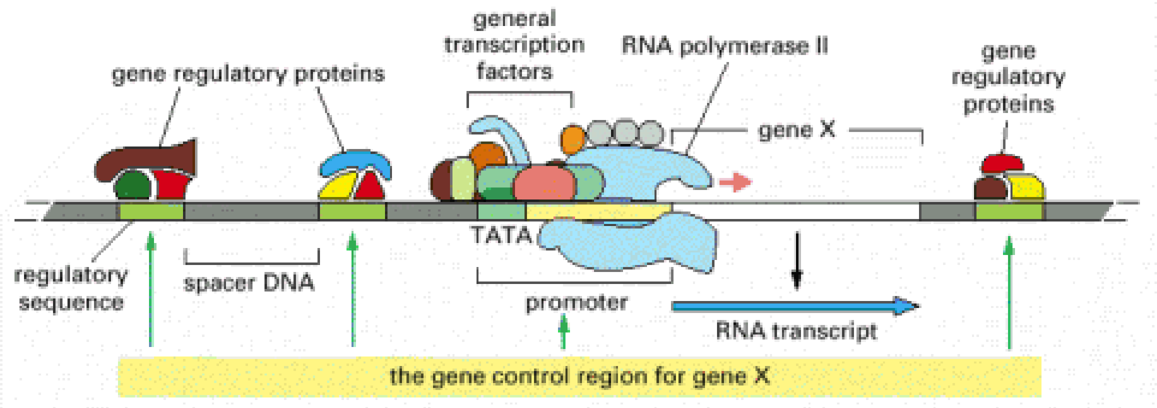


Figure 17. Gene control region of a typical eukaryotic gene.
The promoter is the DNA sequence where the general transcription factors and the polymerase assemble. The *regulatory sequences* serve as binding sites for gene regulatory proteins, whose presence on the DNA influences the rate of transcription initiation. These sequences can be located adjacent to the promoter, far upstream of it, or even within introns or downstream of the gene. DNA looping is thought to allow gene regulatory proteins bound at any of these positions to interact with the proteins that assemble at the promoter. Whereas the general transcription factors that assemble at the promoter are similar for all polymerase II transcribed genes, the gene regulatory proteins and the locations of their binding sites relative to the promoter are different for each gene (Alberts, 2002).

4.1.2 Posttranscriptional controls

In the following section, we consider the variety of posttranscriptional regulation in temporal order, according to the sequence of events that might be experienced by an RNA molecule after its transcription has begun (Figure 18).

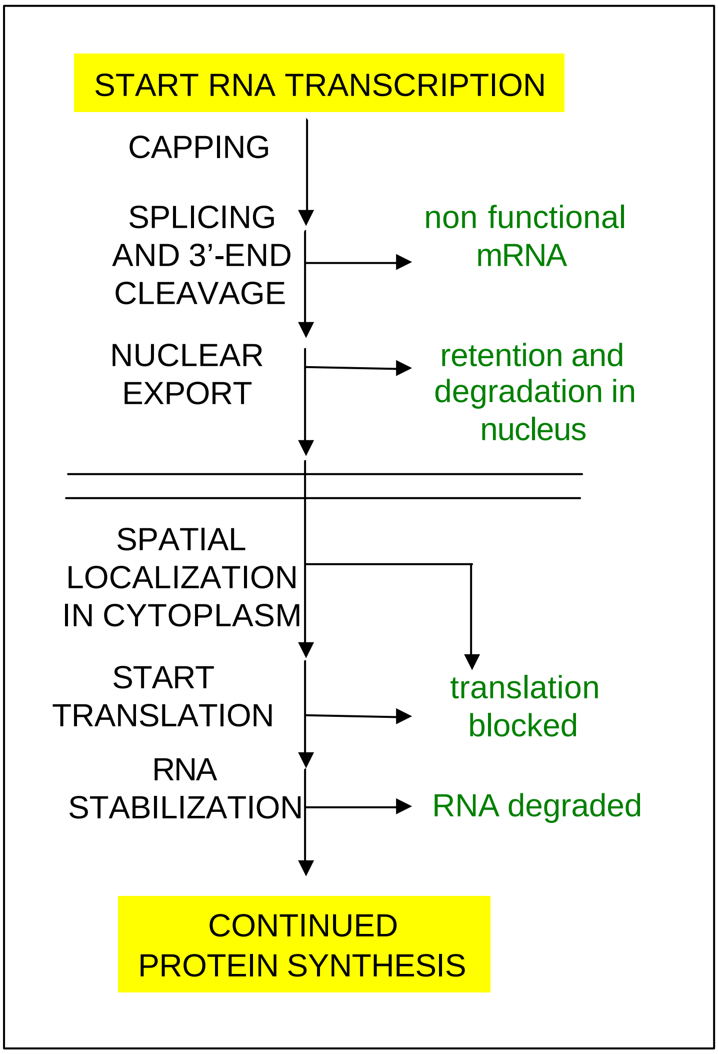


Figure 18. Possible posttranscriptional controls on gene expression.
The possible effects when post-transcriptional controls are disturbed are shown in green.

4.1.2.1 Capping

5' Capping of messenger RNA (mRNA) is essential for cell growth (Shibagaki *et al.*, 1992). It occurs selectively on RNA polymerase II (Pol II) nascent transcripts (20-25 nucleotides in length) and is the first of many pre-mRNA processing events in the nucleus. Caps not only identify gene transcription start sites but also have important effects on mRNA maturation, translation and stability. The 5' methyl cap signals the 5' end of eukaryotic mRNAs and this landmark helps the cell to distinguish mRNAs from the other types of RNA molecules present in the cell. In the nucleus, the cap binds a protein complex called CBC (Cap-Binding Complex) that helps the RNA to be properly processed and exported. The 5’ methyl cap also has an important role in the translation of mRNAs in the cytosol (Shatkin and Manley, 2000).

4.1.2.2 Splicing

Alternative RNA splicing can produce different forms of a protein from the same gene. In addition to switching from the production of a functional protein to the production of a non-functional one, the regulation of RNA splicing can generate different versions of a protein in different cell types, according to the needs of the cell.

RNA splicing can be regulated either negatively, by a regulatory molecule that prevents the splicing machinery from gaining access to a particular splice site on the RNA, or positively, by a regulatory molecule that helps direct the splicing machinery to an otherwise overlooked splice site (Graveley, 2001).

4.1.2.3 RNA transport and localization

The export of RNA molecules from the nucleus is delayed until processing has been completed. Therefore, any mechanism that prevents the completion of RNA splicing on a particular molecule could in principle block the exit of that RNA from the nucleus. Once the newly synthesized eukaryotic mRNA molecule has passed through a nuclear pore and entered the cytosol, it is typically met by ribosomes, which translate it into polypeptide chain. Nevertheless, some mRNAs can be directed to specific intracellular locations before translation begins. The signals that direct mRNA localization are typically located in the 3’ UnTranslated Region (UTR) of the mRNA molecule. The mRNA that encodes actin, for example, is localized to the actin-filament-rich cell cortex in mammalian fibroblasts by means of a 3’ UTR signal (Jansen, 2001).

4.1.2.4 Proteins that bind to the 5’ and 3’ untranslated regions of mRNAs mediate negative translational control

Once an mRNA has been synthesized, one of the most common ways of regulating the levels of its protein product is by controlling the step in which translation is initiated. The selection of an AUG codon as a translation start site is largely determined by its proximity to the cap at the 5’ end of the mRNA molecule, which is the site at which the small ribosomal subunits binds to the mRNA and begins scanning for an initiating AUG codon. Despite the differences in translation initiation, they also utilize translational repressors. Some bind to 5' end of the mRNA and thereby inhibit translational initiation. Others recognize nucleotide sequences in the 3' UTR of specific mRNAs and decrease translation initiation by interfering with the communication between the 5' cap and 3' poly-A tail, which is required for efficient translation.

4.1.2.5 Gene expression can be controlled by a change in mRNA stability

Two major degradation pathways exist for eukaryotic mRNAs : the deadenylation-dependent decay and the deadenylation-independent decay (Garneau *et al.*, 2007).

Once in the cytosol, the poly-A (poly adenylation) tails (which average about 200 adenosine nucleotides) in length) are gradually shortened by an exonuclease that chews away the tail in 3' to 5' direction. Once the critical threshold of tail shortening has been reached (~ 30 adenosine nucleotides remaining), the 5' cap is removed (a process called "decapping"), and the RNA is rapidly degraded. In addition, many mRNAs carry in their 3' UTR sequences binding sites for specific proteins that increase or decrease the rate of poly-A tail shortening.

Deadenylation-independent decay begins with the action of specific endonucleases, which simply cleave the poly-A tail from the rest of the mRNA in one step. The mRNAs that are degraded in this way carry specific nucleotide sequences, typically in their 3' UTR, that serve as recognition sequences for the endonucleases.

Nonsense-mediated mRNA decay

The eukaryotic cell has an additional mechanism, called nonsense-mediated mRNA decay, which eliminates certain types of aberrant mRNAs before they can be efficiently translated into protein. This mechanism was discovered when mRNAs contain misplaced in-frame translation stop codon (UAA, UAG, UGA). These stop codons can arise either from mutation or from incomplete splicing. This mRNA surveillance system therefore prevents the synthesis of abnormally truncated proteins which can be dangerous to the cell. Nonsense-mediated decay may have been especially important in evolution, allowing eukaryotic cells to more easily explore new genes formed by DNA rearrangements, mutations, or alternative patterns of splicing by selecting only those mRNAs for translation that produce a full-length protein (Hentze and Kulozik, 1999).

4.1.2.6 RNA interference is used by cells to silence gene expression

RNAi is a sequence-specific, post-transcriptional, gene-silencing process (Sharp, 1999), (Novina and Sharp, 2004), (Meister and Tuschl, 2004) that is mediated by double-stranded RNA (dsRNA) molecules. The effectors of RNAi are small interfering RNAs (siRNAs) that are processed from longer precursors by a ribonuclease known as DICER. One strand of the siRNA functions as a template for the RNA-induced silencing complex (RISC) to pair to, and cleave, a complementary mRNA. Cleaved mRNAs are then rapidly degraded. Recently, RNA interference (RNAi) has been used by the cell biologist to perturb gene function. This provides a high-throughput approach that can be applied on the cell or organism scale (Hannon and Rossi, 2004).

4.2 Tools for mining the transcriptome

Essentially every cell in an organism is, at any given time point, transcribing thousands of its genes in various quantities. As described in the previous section, the amount of a messenger RNA transcript is tightly regulated and it is critical, both from basic science and clinical perspectives, to be able to accurately quantify the levels of different transcripts.

In this section both gene-by-gene methods and global methods for quantification of messenger RNA levels are described. The focus will be on microarray technology, which is described in the global methods subsection.

4.2.1 Gene-by-Gene methods

Northern blot provides a quantification and size determination of a transcript in a complex mixture by first separating the transcripts by denaturing agarose gel electrophoresis, followed by a transfer to a membrane strip and hybridisation with a labelled probe (Alwine *et al.*, 1977). However, the method is sensitive to RNA degradation and lacks a wide dynamic range. Labelling of the probes is commonly achieved using either radioisotopes or biotin. Typically, DNA probes up to several hundred base pairs (bp) are used, but recently locked nucleic acid (a nucleic acid analogue) has been demonstrated to achieve a 10-fold improvement in sensitivity (Valoczi *et al.*, 2004).

Quantitative real-time reverse-transcription PCR (qRT-PCR) provides superior sensitivity for analysis of gene expression levels compared to other methods; analysis of even single cells is possible (reviewed in (Wong and Medrano, 2005)). First, a complex mixture of total RNA is converted to cDNA using reverse transcriptase with either random or gene-specific priming. Next, a 100-200 bp fragment is amplified and the accumulation of product measured after each cycle using a fluorophore that either specifically targets the amplicon or any double-stranded DNA. During the exponential phase of the amplification each PCR cycle doubles the amount of product and in log2 scale this corresponds to a linear increase. Extrapolation of the linear increase to the level of background provides an estimate of the initial starting amount of mRNA. Use of qRT-PCR has many advantages, making it the method-of-choice for high-accuracy (but low throughput) gene expression analysis:

1) It can achieve single-copy detection (Palmer *et al.*, 2003),

2) It can be carried out in one step

3) It has low coefficients of variation facilitating detection of small differences between samples (Gentle *et al.*, 2001)

4) Design of specific amplicons allows for discrimination between similar transcripts, such as gene family members.

4.2.2 Global methods

Global methods allow for a nearly-complete analysis of the transcriptome.

4.2.2.1 Expressed Sequence Tag (EST) sequencing

An Expressed Sequence Tag is a short (200 to 500 nucleotides long) sequence of transcribed spliced nucleotide sequence. ESTs have proven to be a powerful, inexpensive and rapid approach to identify new genes that are preferentially expressed in certain tissue or cell types (Adams *et al.*, 1991; Adams *et al.*, 1995; Mu *et al.*, 2001). ESTs have also been used for physical mapping, as has been demonstrated in the development of human and mouse gene maps (Hayes *et al.*, 1996; McCarthy *et al.*, 1997; Deloukas *et al.*, 1998). The availability of large numbers of ESTs for individual organisms provides the basis for development of expression arrays.

Currently there are over 45 million ESTs in the NCBI public collection (**http://www.ncbi.nlm.nih.gov/dbEST/**) (Table 2).With many large-scale EST sequencing projects newly initiated or in progress; the number of ESTs in the public domain will continue to increase in the coming years.

Species	**Number of ESTs**
Homo sapiens	8,133,717
Mus musculus	4,850,243
Bos taurus	1,496,055
Sus scrofa	1,470,315
Danio rerio	1,358,222
Rattus norvegicus	889,896
Gallus gallus	599,330
Drosophila melanogaster	541,595

Table 2. Number of EST sequences per organisms in October 2007

4.2.2.2 Serial Analysis of Gene Expression (SAGE)

Serial analysis of gene expression (SAGE) is a method that allows the quantitative and simultaneous analysis of large number of transcripts (Velculescu *et al.*, 1995).

Three principles underlie the SAGE methodology:

1) A short sequence tag (10-14bp) contains sufficient information to uniquely identify one transcript, given that the tag is obtained from a unique position within each transcript.

2) Sequence tags can be linked together to form long serial molecules that can be cloned and sequenced;

3) Quantification of the number of times a particular tag is observed provides the expression level of the corresponding transcript.

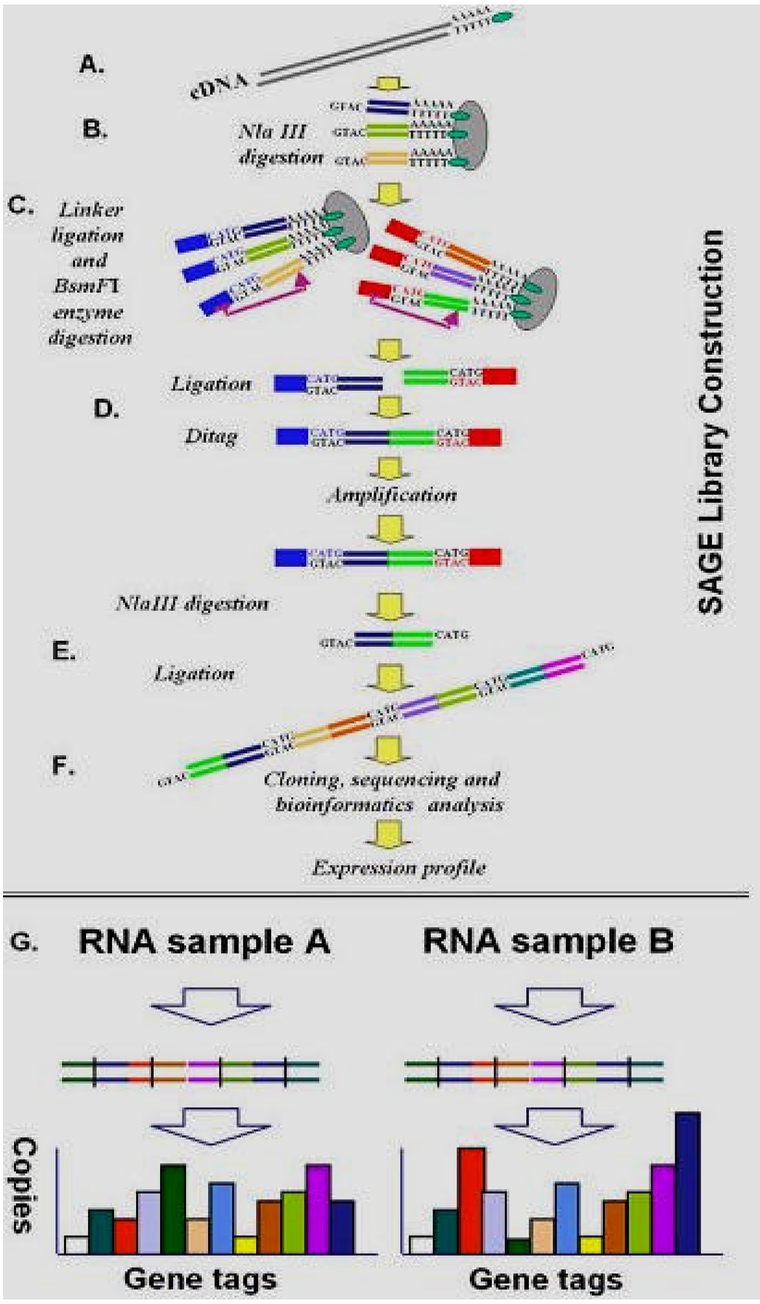


Figure 19. Serial analysis of gene expression (SAGE) library construction. Steps in SAGE profiling include: (A) An RNA population is reverse transcribed to cDNAs using oligo-T primers attached to magnetic beads. (B) cDNAs are collected and digested with the restriction endonuclease Nla III. (C) Linkers containing sequence recognized by BsmF I are ligated to the digested cDNAs. Sequence tags are released from the beads by BsmF I digestion (BsmF I cuts at a fixed distance downstream from its recognition site). (D) Released DNA tags are ligated together to form ditags. (E) Ditags are amplified and then digested with Nla III to remove the linkers. (F) Ditags are ligated together to form a concatemer which is then cloned into a plasmid vector to generate a SAGE library. The identity and abundance of tags is deduced from DNA sequence analysis of plasmid clones of concatenated ditags. (G) Relative abundance of gene expression – between genes within the same RNA population or between samples – is deduced by counting sequence tags, adapted from (Garnis *et al.*, 2004).

The SAGE method (Figure 19) has demonstrated its effectiveness at cataloguing large quantities of expressed genes in cells or tissues from a variety of physiological, developmental and pathological states (Madden *et al.*, 1997; Zhang *et al.*, 2006).

4.2.2.3 Microarray based methods

Microarray-based technology provides investigators with the ability to measure the expression profile of thousands of genes in a single experiment. The use of miniaturized microarrays for gene expression profiling was first published in 1995 (Schena *et al.*, 1995). The basic principle underlying microarray technology is that complementary nucleic acids will hybridize. This is also the basis for traditional gene expression analyses, such as Southern and Northern blotting. Hybridization provides excellent selectivity of complementary stranded nucleic acids, with high sensitivity and specificity. In the traditional techniques, in which radioactive labeling materials are usually used, the simultaneous hybridization of test and reference samples is impossible. In microarray-based technologies, the solid surface, such as a glass slide, contains hundreds to thousands of immobilized DNA (targets) spots which can be simultaneously hybridized with two samples (probes) labeled with different fluorescent dyes. After hybridization, the fluorescent signals of two probes bound to individual spots are detected with a confocal laser scanner. The separately scanned images from each of the two probes are then combined and pseudo-colored by means of computer software and the expression ratios of two probes are calculated. Based on these systematic procedures, microarrays make large-scale gene expression monitoring possible in a parallel fashion.

Types of microarrays

The two basic types of microarrays are spotted arrays and oligonucleotide arrays.

In **spotted microarrays**, the probes are oligonucleotides (60-75bp), cDNA or small fragments of PCR products resulting from mRNAs and are spotted onto the microarray surface. This type of array is typically hybridized with cDNA from two samples to be compared (e.g. diseased tissue versus healthy tissue) that are labeled with two different fluorophores. Fluorescent dyes commonly used for cDNA labelling include Cyanine3 (Cy3), which has a fluorescence emission wavelength of 570 nm (corresponding to the green part of the light spectrum), and Cyanine5 (Cy5) with a fluorescence emission wavelength of 670 nm (corresponding to the red part of the light spectrum). The two Cy-labelled cDNA samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores after excitation with a laser beam of a defined wavelength (Figure 20). Relative intensities of each fluorophore may then be used in ratio-based analysis to identify up-regulated and down-regulated genes. Absolute levels of gene expression cannot be determined in the two-colour array.

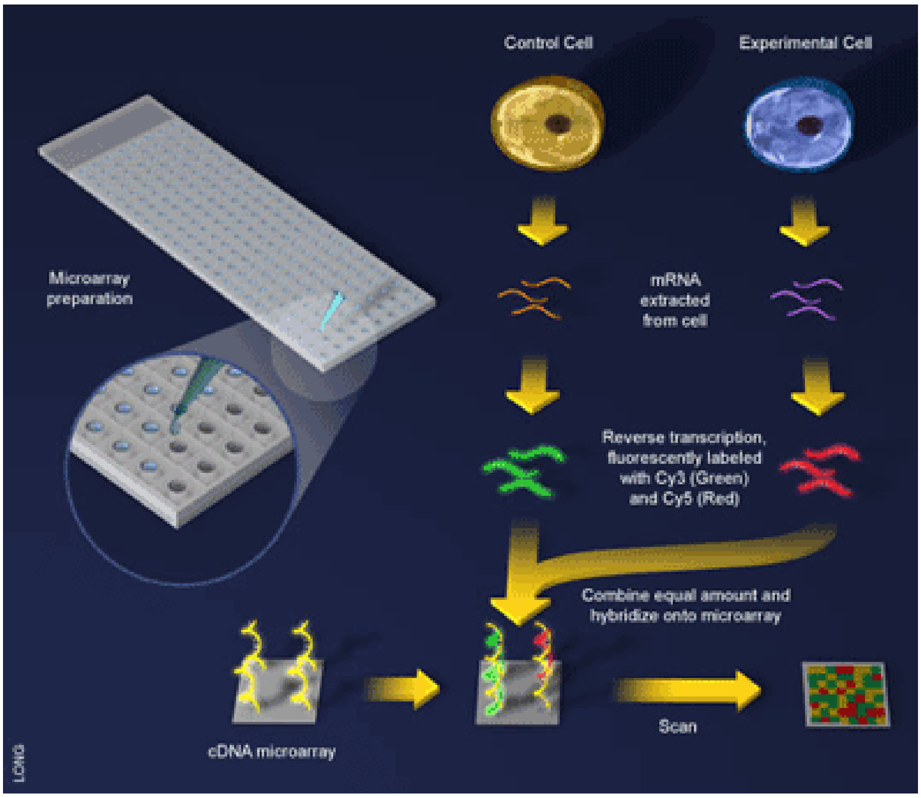

Control Cell
Experimental Cell
Microarray preparation
mRNA extracted from cell
Reverse transcription, fluorescently labeled with Cy3 (Green) and Cy5 (Red)
Combine equal amount and hybridize onto microarray
cDNA microarray
Scan
LONG

Figure 20. An example of a cDNA Microarray Experiment.
RNA is extracted from two different samples and converted into complementary DNA (cDNA), during which the DNA is labeled with fluorescent compounds. The two samples are then mixed together for comparison and hybridized to the array. Differences in gene expression are revealed by fluorescent patterns on the array. Red indicates higher expression, green indicates lower expression and yellow indicates equal expression of gene in experiment cell compared to control cell. The figure was obtained from http://www.scq.ubc.ca.

In **oligonucleotide microarrays**, the probes are designed to match parts of the sequence of known or predicted mRNAs. There are commercially available designs that cover complete genomes from companies such as Affymetrix, Agilent (expect of dual mode gene expression array) or GE Healthcare. These microarrays give estimations of the absolute value of gene expression and therefore the comparison of two conditions requires the use of two separate microarrays. Long Oligonucleotide Arrays are composed of 60-mers or 50-mers and are produced by ink-jet printing on a silica substrate. Short Oligonucleotide Arrays are composed of 25-mer or 30-mer and are produced by photolithographic synthesis (Affymetrix) on a silica substrate or piezoelectric deposition (GE Healthcare) on an acrylamide matrix. Oligonucleotide microarrays often contain control probes designed to hybridize with RNA spike-ins. The degree of hybridization between the spike-ins and the control probes is used to normalize the hybridization measurements for the target probes.

Genechip arrays

The GeneChip (Affymetrix) high-density oligonucleotide arrays (Figure 21) are fabricated by using in-situ synthesis of short oligonucleotide sequences on a small glass chip using light directed synthesis. This technique allows for the precise construction of a highly ordered matrix of DNA oligomers on the chip.

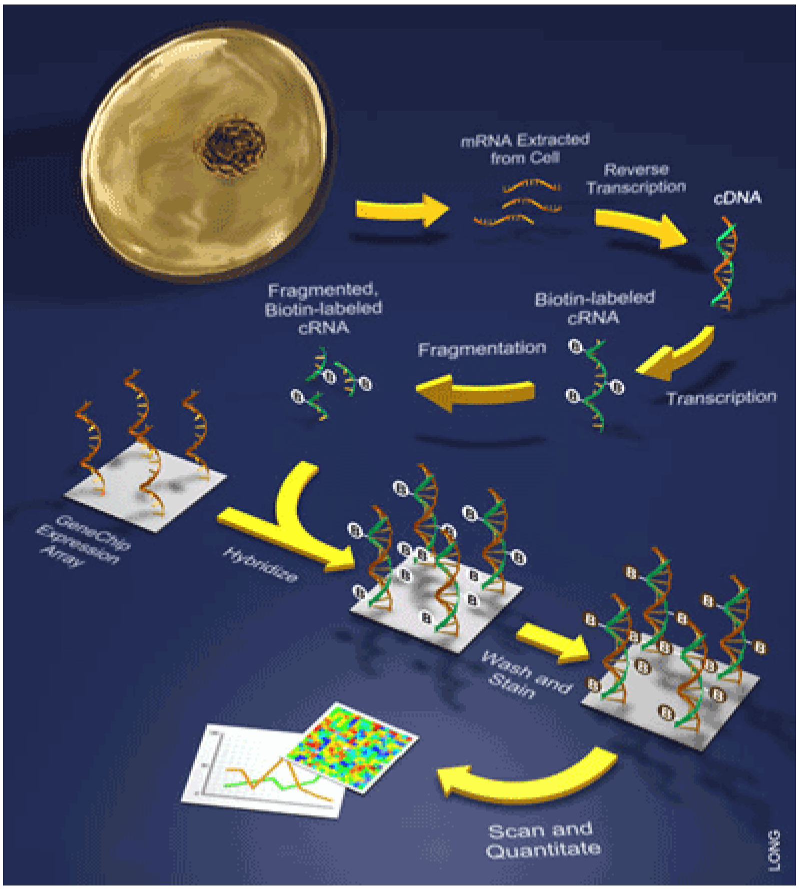


Figure 21. The Use of Oligonucleotide Arrays.
mRNA is extracted from cells and amplified through a process that labels the RNA for analysis. The sample is then applied to an array and any bound RNA stained. The figure was obtained from http://www.scq.ubc.ca.

In the GeneChip system a known RNA transcript or potentially expressed sequence is represented on the chip by 11-20 unique oligomeric probes, each 25 bases in length. The group of oligomeric probes corresponding to a given gene or small group of highly similar genes is known as the probe set and generally spans a region of about 600 bases, known as the target sequence. Many copies of each oligomeric probe are synthesized in discrete features (or cells) on the GeneChip array. In addition, for each oligomer on the array there is a matched oligomer, synthesized in an adjacent cell that is identical with the exception of a mismatched base at the central position (i.e. base 13). These are designated Perfect Match (PM) and Mismatch (MM) probes, respectively. The MM probes serves as a control for non-specific hybridization (Figure 22).

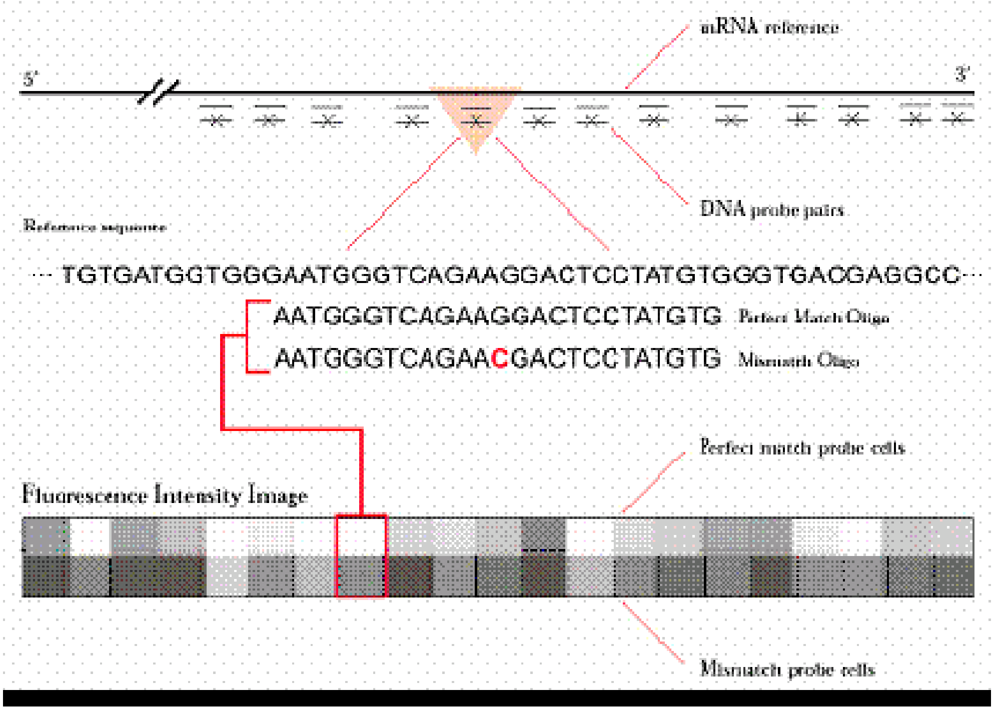

mRNA reference
5'
3'
DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC···
AATGGGTCAGAAGGACTCCTATGTG
Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTG
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 22. Illustrates the relationship between perfect and mismatch probe sequences.
The Affymetrix GeneChip technology. The presence of messenger RNA (mRNA) is detected by a series of probe pairs that differ in only one nucleotide. Hybridization of fluorescent mRNA to these probes pairs on the chip is detected by laser scanning of the chip surface. A probe set = 11-20 PM, MM pairs. Figure adapted from (http://keck.med.yale.edu/affymetrix/technology.htm).

The data obtained after scanning the hybridized Affymetrix array is usually in the form of .CEL files. The CEL file contains only intensity information for a given probe on an array.

4.3 Applications of microarrays

Microarrays have been used successfully in various research areas including: sequencing, single nucleotide polymorphism (SNP) detection, characterization of protein-DNA interactions, mRNA profiling, and many more.

Applications of microarrays in the biosciences include: gene expression studies, disease diagnosis, pharmacogenomics, drug screening, pathogen detection, and genotyping, summarized in (Table 3). In addition, tissue microarrays can be used for quality assurance for immunohistochemical and in situ hybridization procedures (van de Rijn and Gilks, 2004).

Studies	**Probe (Printed)**	**Labelled target**
Expression studies	cDNA or oligonucleotides	Total RNA or mRNA
Comparative genomic hybridisation (CGH)	cDNA, BACS, oligonuc leotide	PCR products/genomic DNA
Mutation detection (gene specific)	Oligonucleotides, primer extension	PCR amplification of the gene to be tested
SNP detection	Oligonucleotides, primer extension	PCR products/genomic DNA
Chromatin immunoprecipitation (ChIP)	cDNA, CpG island clones, genomic sequences, BACS, Oligonucleotides	Sonicated and immunoprecipitated genomic DNA/genomic DNA

Table 3. Applications of Microarray technology.

Chapter 5. Microarray Data Analysis

The key steps required for conversion of microarray data into meaningful biological knowledge can be divided into: identification of significantly regulated genes, identification of global patterns of gene expression and determination of the biological meaning of both individual genes and groups of genes (Draghici, 2002).

5.1 Experimental design

The type of experiment will affect the downstream statistical analysis of the data. The simplest experiments will involve a comparison between two conditions, e.g. control and test will require a two-group statistical test such as a t test (Gayen, 1949). A multiple condition experiment (such as a time course) requires a test that can be used for more than two samples, such as analysis of variance (ANOVA). The type of ANOVA needed will strongly depend on the experimental design, in particular how many factors are being examined. Examples of factors that might be examined in typical microarray experiments include disease states, strains of model organisms, drug treatments, or times after treatment. If one factor is being studied, e.g. time after treatment in a time series, a one-way ANOVA may be used. If multiple factors are being studied, then a two-way ANOVA is needed (Pavlidis, 2003).

5.1.1 Replicates

The use of replicates helps to distinguish between genes that are truly differentially expressed and those affected only by noise. Averaging over replicates minimizes the effects of chance variation and allows the extent of experimental variation to be estimated. Comparison tests use this estimate of variability within the replicates to assign a confidence level as to whether the gene is differentially expressed. Two types of replicates (Figure 23) are generally considered for microarray experiments: technical and biological. Technical replicates deal with technical variation and an example is hybridizing the same sample different times but with the same type of arrays. Biological replicates are used to deal with biological variability that may be due to genetic variability or environmental effects. An example of biological replicates is the extraction of RNA from retina from three genetically identical mice and the hybridization of each sample to its own array. Generally three replicates should be considered a minimum, but studies suggest that four or five biological replicates would be better (Pavlidis *et al.*, 2003). The exact number of replicates required depends on the variability inherent in the system being studied, with studies using patient data usually requires more than for studies using animal models or cell lines, since in general sample from patients show more variation than animal models or cell lines. Regardless of the system being studied, it is important to realize that rigorous statistical inferences cannot be made with a sample size of one and, in general, the more replicates, the stronger the inferences that can be made from the data.

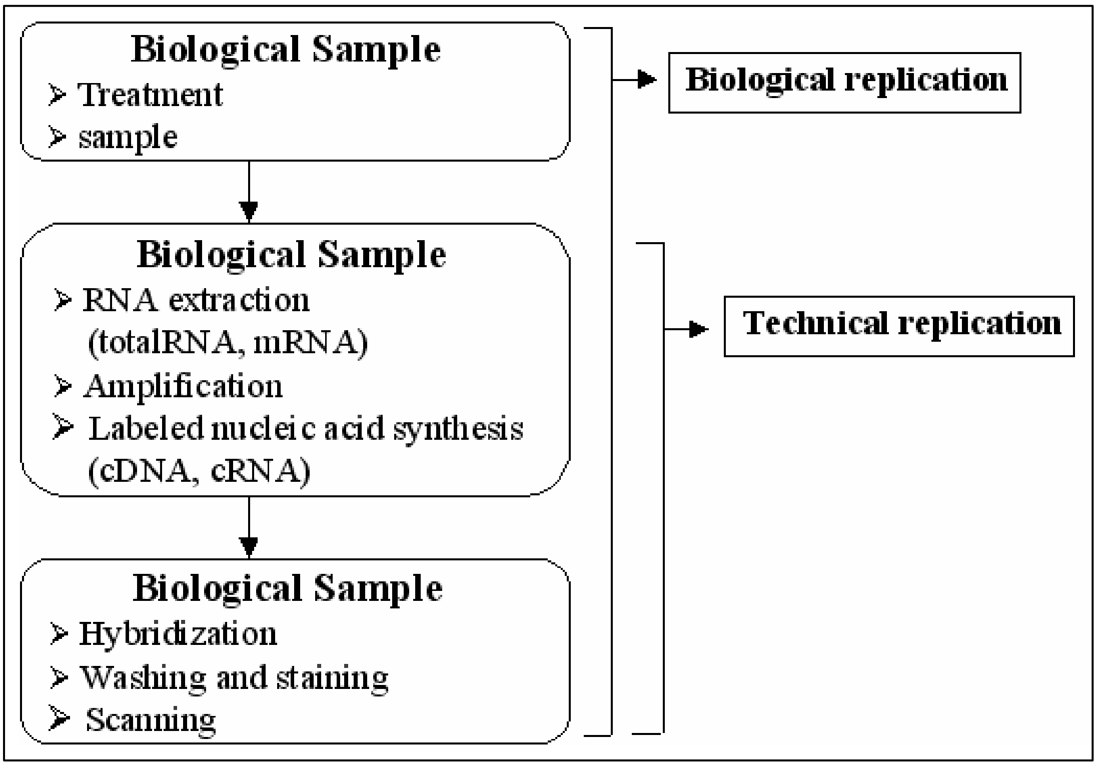


Figure 23. Origin of different types of replication.

5.2 Data preprocessing and normalization

The need for data preprocessing and normalization comes from the fact that, the spot intensities directly reflects mRNA levels, but these intensities may also depend on peculiarities of print tips, particular PCR reactions, integration efficiency of a dye and hybridization specific effects. Also, measurements from different hybridizations may occupy different scales. So, in order to compare them they need to be normalized, otherwise one could deem genes differentially expressed where only the hybridizations behaved differently. Additionally, the variance in the data tends to depend on the absolute intensity of the data. This may also lead to false biological conclusions and should be remedied by a normalization method.

There are four main steps that need to be carried out during a microarray preprocessing procedure:

- *Background correction*: (to adjust for hybridization effects that are not associated with the specific interaction of probe with target DNA),
- *Normalization* (adjust chips to a common baseline so that they are comparable among each other),
- *Perfect Match (PM)/MisMatch (MM)* correction (adjust for non-specific hybridization of target to probe) in case of oligonucleotide arrays
- *Summarization* (summarize different probes of a probeset to yield a single expression value per gene).

5.2.1 Different algorithms used in preprocessing of oligonucleotide array data

Different normalization methods to yield a single expression value for each gene are DNA-Chip analyzer (dChip) which uses Model Based Expression Index (MBEI) algorithm of Li and Wong (Li and Wong, 2001), the MAS 5.0 Statistical algorithm from Affymetrix, Robust Multichip Average (RMA) proposed in Irizarry *et al.* (Irizarry *et al.*, 2003b) and sequence based method of background adjustment GeneChip RMA GCRMA (Wu and Irizarry, 2005). The Bioconductor package (http://bioconductor.org) (an open source and open development software project for the analysis and comprehension of genomic data) provides implementations for a number of methods that combine these four algorithms.

MAS5.0 takes a robust average of log (PM - MM) using one step Tukey’s biweight estimate, where outliers are penalized with low weights. Then, it normalizes arrays by scaling each array so that all arrays have the same mean.

dChip (http://www.dchip.org/) chooses an array (the default is the one with median overall intensity) to normalize other arrays against the array at the probe intensity level. Normalization is done by determining the normalization curve with a subset of probes, or invariant probe sets. The resulting signals, or Model-based expression indexes, are either the weighted average of PM/MM differences (PM/MM model) or background-adjusted PM values (PM-only model) of selected probes estimated using a multiplicative model. The model-fitting and outlier-detection are iterated until the set of array, probe and single outliers is stabilized, and then those outliers can be excluded or imputed.

RMA (Robust Multichip Average) is another probe set algorithm (or summary measure) that takes a Robust Multi-array Average (RMA) of background adjusted, normalized (quantile normalization) and log transformed PM only values. A robust procedure called 'median polish' is used to estimate parameters of an additive model (Irizarry *et al.,* 2003a).

GCRMA uses GC content of probes with RMA. In general, RMA is best for projects with less confounding noise, where it provides high sensitivity at low replicates.

On the other hand, dChip appears to perform best in noisier projects, with fewer false positives (Seo *et al.,* 2006).

We have used both dChip, RMA and MAS 5.0 (for some experiment) methods to obtain signal intensity for the individual probesets for all the experiments in **RETINOBASE**, the web database, data mining and analysis platform for gene expression data on retina are described in Chapter 10.

5.3 Biological analysis through differential expression

The task of obtaining differentially expressed genes from an experiment can be divided into the following steps:

(1) *Ranking*: genes are ranked according to the evidence of differential expression between the control and test sample.

(2) *Assigning significance*: a statistical significance is assigned to each gene.

(3) *Cut-off value*: to obtain a limited number of differentially expressed genes, a cut-off value for the statistical significance needs to be determined.

The quantification of gene expression differences depends critically on the experimental setting. The first distinction depends on whether replicates are available or whether the measurement has been made only once. In the absence of replicates, the options are very limited, but it is possible to use a one sample student test. Availability of repetitions means that many more statistical procedures are applicable (Figure 24).

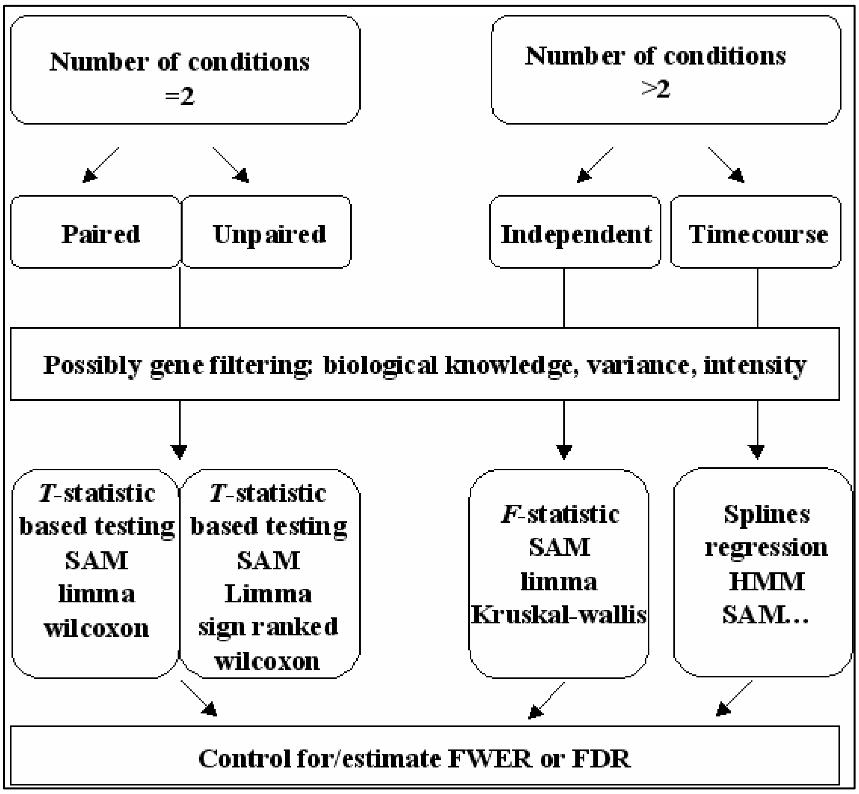

Number of conditions =2
Number of conditions >2
Paired
Unpaired
Independent
Timecourse
Possibly gene filtering: biological knowledge, variance, intensity
T-statistic based testing SAM limma wilcoxon
T-statistic based testing SAM Limma sign ranked wilcoxon
F-statistic SAM limma Kruskal-wallis
Splines regression HMM SAM...
Control for/estimate FWER or FDR

Figure 24. Methods for quantification of differential gene expression in replicated experiments.
Basic considerations for an appropriate testing procedure aimed at quantifying differential expression. Figure adapted from (Steinhoff and Vingron, 2006).

The most suitable procedure depends on (i) the number of conditions that are compared and (ii) one also distinguishes between independent and dependent settings.

5.3.1 Two-conditional setting and independent multi-conditional setting

In a two-condition case (typically, test versus control conditions), one considers either a paired/unpaired situation. An example of a paired situation is gene expression measurements of one cell line before and after chemical treatment. An unpaired experiment involves independent samples, such as the comparison of a healthy group with a diseased one. An example for the independent multi-conditional setting is finding differentially expressed genes comparing multiple groups of disease stages. Most commonly, multi-conditional experiments are time courses.

As previously discussed, the availability of replicates enables to rank genes according to their associated t-statistic for each gene: $t = m/(std/\sqrt{n})$, where *m* is the difference of means across replicates, *std*, the within groups standard deviation and *n*, the number of genes considered for testing. *F*-scores are the straightforward generalization of t-scores in the multi-conditional case. There will be a problem when genes with low intensity show almost no changes between conditions. This might show high *t*-scores. A possible solution might be to enlarge these variances artificially. Low variances can be corrected by proposing an enlarging factor e.g. Fudge factor (variables whose purpose to force a calculated result) (Tusher *et al.*, 2001). For example, the approach by Tusher et al. (Tusher *et al.*, 2001) is implemented in the computational tool called Significance Analysis of Microarrays (SAM).

Several linear model approaches for ranking gene expression differences have been introduced (Kerr *et al.*, 2000), (Smyth, 2004), (Thomas *et al.*, 2001), (Lin *et al.*, 2004). Kerr et al. (Kerr *et al.*, 2000) used ANOVA models for an integrated procedure of normalization and detection of differentially expressed genes. They assumed a linear model of specific effects for log intensities of all genes. These effects might be from slide printing, possible variations in sample treatment, gene effects and their respective interactions. Smyth (Smyth, 2004) proposed a modified t-statistic that is proportional to the t-statistic with sample variance offset as used in (Efron B, 2001; Tusher *et al.*, 2001; Broberg, 2003). The approach can be generalized for the multi-conditional case. It has been implemented in the Bioconductor package limma (Linear Models for MicroArray data) (Smyth and Speed, 2003; Smyth, 2004). Using this package, experimental setting, duplicate spots and quality weights can also be considered. The moderated t-statistic is calculated, genes are ranked with respect to the resulting scores and *p*-values can be assigned.

Furthermore, a number of rank-based approaches (thus, non-parametric) have been developed. These are based on a Wilcoxon rank sum test or permutation t-test. While t-test and F-test based methods assume that the intensity measurements of normalized ratios are normally distributed, rank-based approaches do not do so. Instead of considering numerical values, Wilcoxon rank sum tests use ranks. This is a more robust approach, although frequently with lower power, because one loses information by switching from the numerical to the rank scale. In the multi-conditional case, the Kruskal–Wallis test is the straightforward generalization of the Wilcoxon test.

Dependent multiconditional setting

In time course experiments, each time point represents one conditional group. All experiments corresponding to one time point build up one conditional group. The essential difference compared with independent cases is the linear order of states. Statistically, each time conditional state, e.g. each time point, is dependent on all the others. This fact requires new statistical procedures. The original form of SAM (Tusher *et al.*, 2001) has been generalized to time course experimental settings.

5.3.2 Cut-off and multiple testing

After ranking the genes according to a statistical procedure, one has to find a cut-off above which biologically meaningful information is expected. The *p*-value measures consistency by calculating the probability of observing the results from your sample of data or a sample with results more extreme, assuming the null hypothesis is false; the smaller the *p*-value, the greater the consistency. Traditionally, researchers will reject a hypothesis if the *p*-value is less than 0.05 and assume all genes showing a lower *p*-value to be biologically significant. Performing many tests at a time, however, increases the problem of falsely significant genes. Roughly speaking, when performing 10, 000 tests one expects 5% of the genes to show a *p*-value of less than 0.05 just due to chance. There are a number of multiple testing approaches to overcome this problem. One possibility to lower the problem is to reduce the number of statistical tests by filtering steps. Thus, we have to find a criterion for limiting the number of testing procedures. This might be either external biological knowledge or variance across conditions. Thus, the set of intensities can be reduced by neglecting genes from which we do not expect any biological information. Alternatively, one could use only those genes that show a certain minimal amount of variance over all conditional states or apply intensity-based filtering, e.g. neglecting very lowly expressed genes (Steinhoff and Vingron, 2006).

For the oligo-array experimental setting, the statistical confidence that the ensembled perfect match signals from a probe set are significantly above background is derived as a “detection *p*-value”. If, on average, the perfect match probes show greater signal intensity than the corresponding mismatch, then the detection *p*-value improves, to a threshold to where a “present call” is assigned (e.g. the target transcript is likely “present” in the sample tested). Whereas poor signal from the perfect match, and increasing signal from the mismatch, causes the detection *p*-value to worsen, leading to an “absent call”. The “present call” is thus a statistically derived threshold reflecting a relative confidence that the desired mRNA is indeed present in the RNA sample being tested at a level significantly above background hybridization. An "absent call" suggests that the target mRNA is not present in the sample, or that there is non-target mRNAs binding to the probe set (non-desired cross-hybridization), or both. Data can be filtered based on these calls.

Given a type I error rate (i.e. a false positive rate) controlling for multiple testing means correcting *p*-values such that the given error rate can be guaranteed for all tests. Methods can be divided into those that control the Family Wise Error Rate (FWER) or the False Discovery Rate (FDR). The probability of at least one type I error within the significant genes is called FWER. The FDR is the expected proportion of type I errors within the rejected hypotheses (Figure 24).

5.4 Data clustering

Clustering techniques applied to gene expression data will partition genes into groups/clusters based on their expression patterns. Genes in the same cluster will have similar expression patterns, while genes in different clusters will have distinct well-separated expression patterns**.**

Clustering methods can be divided into two basic types: hierarchical and partitional clustering. Within each of the types there exists a wealth of subtypes and different algorithms for finding the clusters.

Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters. One of the main differences between the clustering methods lies in the rule used to decide which two small clusters are merged or which large cluster is split. The end result of the algorithm is a tree of clusters called a dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level, a clustering of the data items into disjoint groups is obtained.

Partitional clustering, on the other hand, attempts to directly decompose the data set into a set of disjoint clusters. The criterion function that the clustering algorithm tries to minimize may emphasize the local structure of the data, as by assigning clusters to peaks in the probability density function, or the global structure. Typically the global criteria involve minimizing some measure of dissimilarity in the samples within each cluster, while maximizing the dissimilarity of different clusters. A commonly used partitional clustering method is K-means clustering.

5.4.1 Hierarchical clustering

Hierarchical clustering has the advantage that it is simple and the result can be easily visualized (Eisen and Brown, 1999). It has become one of the most widely used techniques for the analysis of gene-expression data. One of the hierarchical clustering methods is the agglomerative approach in which single expression profiles are joined to form groups, which are further joined until the process has been carried to completion, forming a single hierarchical tree. The process of hierarchical clustering proceeds in a simple manner. First, the pairwise distance matrix is calculated for all of the genes to be clustered. Second, the distance matrix is searched for the two most similar genes or clusters; initially each cluster consists of a single gene. This is the first true stage in the 'clustering' process. If several pairs have the same separation distance, a predetermined rule is used to decide between alternatives. Third, the two selected clusters are merged to produce a new cluster that now contains at least two objects. Fourth, the distances are calculated between this new cluster and all other clusters. There is no need to calculate all distances as only those involving the new cluster have changed. Last, steps 2–4 are repeated until all objects are in one cluster. There are several variations on hierarchical clustering most common once are Single-linkage clustering, Complete linkage clustering, Average linkage clustering that differ in the rules governing how distances are measured between clusters as they are constructed.

Each of these will produce slightly different results, as will any of the algorithms if the distance metric is changed. Typically for gene-expression data, Average linkage clustering gives acceptable results (Quackenbush, 2001).

One potential problem with many hierarchical clustering methods is that, as clusters grow in size, the expression vector that represents the cluster might no longer represent any of the genes in the cluster. Consequently, as clustering progresses, the actual expression patterns of the genes themselves become less relevant. Furthermore, if a bad assignment is made early in the process, it cannot be corrected. An alternative, which can avoid these artifacts, is to use a divisive clustering approach, such as k-means or self-organizing maps, to partition data (either genes or experiments) into groups that have similar expression patterns.

5.4.2 Partitioning clustering: K-means clustering

If there is advanced knowledge about the number of clusters that should be represented in the data, k-means clustering is a good alternative to hierarchical methods (Tavazoie *et al.*, 1999). In k-means clustering, objects are partitioned into a fixed number (k) of clusters, such that the clusters are internally similar but externally dissimilar; no dendograms are produced (but one could use hierarchical techniques on each of the data partitions after they are constructed). The process involved in k-means clustering is conceptually simple, but can be computationally intensive. First, all initial objects are randomly assigned to one of k clusters (where k is specified by the user). Second, an average expression vector is then calculated for each cluster and this is used to compute the distances between clusters. Third, using an iterative method, objects are moved between clusters and intra- and inter-cluster distances are measured with each move. Objects are allowed to remain in the new cluster only if they are closer to it than to their previous cluster. Fourth, after each move, the expression vectors for each cluster are recalculated. Last, the shuffling proceeds until moving any more objects would make the clusters more variable, increasing intra-cluster distances and decreasing inter-cluster dissimilarity.

Some implementations of k-means clustering allow not only the number of clusters, but also seed cases (or genes) for each cluster, to be specified.

5.4.3 Mixture Model clustering

Model-based algorithms are based on the assumption that the whole set of microarray data is a finite mixture of certain type of distributions with different set of parameters. Parameters in such a model define clusters of similar observations. The cluster analysis is performed by estimating these parameters from the data. The model-based algorithm will guarantee finding the optimal clusters if the sample size is sufficiently large. In a Gaussian mixture model approach, similar individual profiles are assumed to have been generated by the common underlying ‘pattern’ represented by a multivariate Gaussian random variable. The confidence in obtained patterns and the confidence in individual assignments to particular clusters are assessed by estimating the confidence in corresponding parameter estimates.

The choice of the optimal number of clusters and the models which fit the data best can be done by using some objective statistical criteria, i.e., Akakike Information

Criterion (**AIC**) (Akaike, 1974) and Bayesian information Criterion (**BIC**) (Schwarz, 1978), whereas for the other algorithms, choosing the 'correct' number of clusters and the best clustering method is still a question open to discussion.

Gaussian finite mixture model (McLachlan and Basford, 1988) based approach to clustering has been used to cluster expression profiles (Yeung *et al.*, 2001). Assuming that the number of mixture components is correctly specified, this approach offers reliable estimates of confidence in assigning individual profiles to particular clusters. In the situation where the number of clusters is not known, this approach relies on ones ability to identify the correct number of mixture components generating the data

The two clustering methods discussed above have been used to cluster all the gene-expression data available in **RETINOBASE**.

5.5 Introduction to meta-analysis

Meta-analysis is a statistical technique for amalgamating, summarizing, and reviewing previous quantitative research (Glass, 1977). By using meta-analysis, a wide variety of questions can be investigated, as long as a reasonable organization of primary research studies exists. Selected parts of the reported results of primary studies are entered into a database, and this "meta-data" is "meta-analyzed", in similar ways to working with other data first descriptively and then inferentially to test certain hypotheses. Meta-analysis provides a systematic overview of quantitative research that has examined a particular question. The danger is that in amalgamating a large set of different studies the construct definitions can become imprecise and the results difficult to interpret meaningfully.

Strengths of Meta-analysis

Meta-analysis imposes a discipline on the process of summing up research finding. It represents findings in a more differentiated and sophisticated manner than conventional analysis. It is capable of finding relationships across studies that are obscured in other approaches and can handle a large numbers of studies. It also protects against over-interpreting differences across studies.

5.5.1 Meta analysis in gene expression studies

Recently, with the advent of large collections of microarray data, obtained through collaborative projects or gene expression repositories, meta-analysis approaches have been applied to the analysis of expression datasets. Rhodes et al. (Rhodes *et al.*, 2002) demonstrated a statistical model for performing meta-analysis of independent microarray datasets. Implementation of this model revealed that four prostate cancer gene expression datasets shared significantly similar results, independent of the method and technology used (i.e., spotted cDNA versus oligonucleotide). This inter-study cross-validation approach generated a cohort of genes that were consistently and significantly dys-regulated in prostate cancer. A recent study by Mulligan et al. applied meta-analysis to a set of microarray studies across mouse lines differing in ethanol-drinking behavior (Mulligan *et al.*, 2006). Genetic studies in mice and humans have implicated an increasing number of genes or genetic intervals as having an influence on drinking behavior. As mentioned above, previous microarray studies on the acute or chronic effects of ethanol have generally involved comparing two conditions (treated and untreated) within a single strain or two phenotypically extreme strains. Mulligan *et al* also studied a two-state model (high versus low ethanol consumption) but did so by comparing 13 different strains in five different experiments from three different laboratories. They also used two different types of microarray platforms (Affymetrix oligonucleotide microarrays and custom spotted cDNA arrays), which gives additional power to the meta-analysis by ensuring that the results are not platform specific.

Chapter 6. Retina

The eye (Figure 25) is a highly specialized organ of photoreception, the process by which light energy from the environment produces the chemical changes in specialized cells in the retina, the rods and cones. These changes result in nerve action potentials, which are subsequently relayed to the optic nerve and then to brain, where the information is processed and consciously appreciated as vision. All the other structure in the eye are secondary to this basic physiological process, although they may be part of the system necessary for focusing and transmitting the light on the retina, for example cornea, lens, iris and ciliary body, or they may be necessary for nourishing and supporting the tissues of the eye, for example the choroid, aqueous outflow system, and lacrimal apparatus.

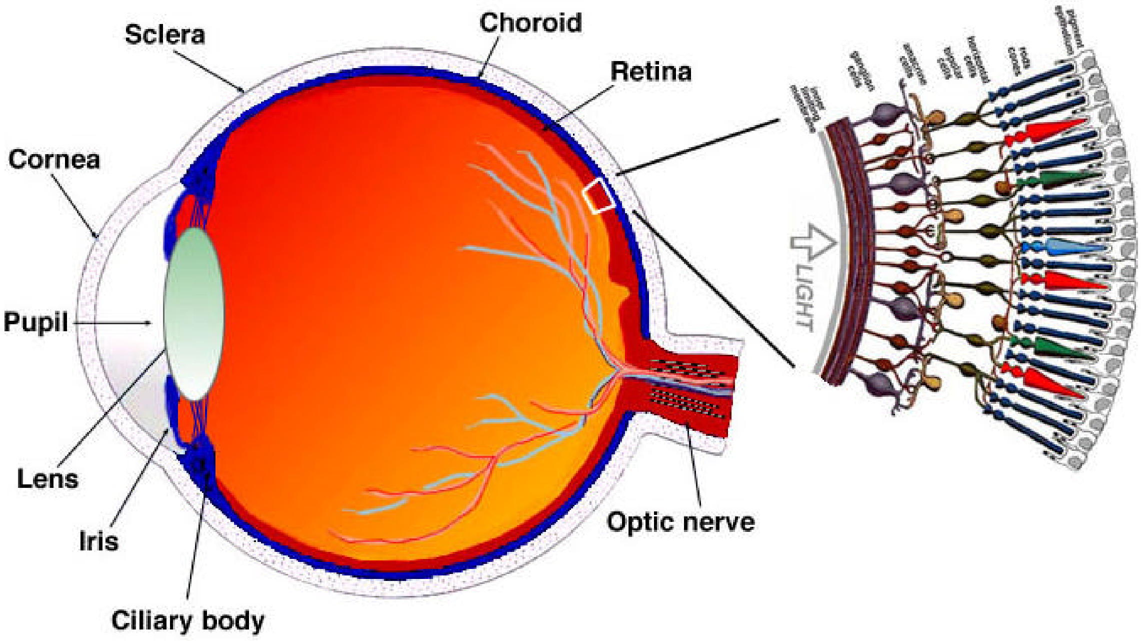


Figure 25. A drawing of a section through the human eye with a schematic enlargement of the retina. Figure taken from (http://webvision.med.utah.edu/sretina.html)

The eye is made up of three basic layers or coats of ten known as tunics (Figure 25). These are the fibrous (corneoscleral) coat, the uvea or uveal tract (composed of choroid, ciliary body, and iris), and the neural layer (retina). The coats surround the contents, namely the lens and transparent media (aqueous humour and vitreous body).

6.1 Retinal development

The retina is a part of the central nervous system (CNS) and a model region of the vertebrae brain to study, because like other regions of the CNS, it derives from the neural tube. The retina is formed during development of the embryo from optic vesicles outpouching from two sides of the developing neural tube. The primordium optic vesicles fold back in upon themselves to form the optic cup with the inside of the cup becoming the retina and the outside remaining a monolayer of epithelium known as retinal pigment epithelium. Initially both walls of the optic cup are one cell thick, but the cells of the inner wall divide to form a neuroepithelial layer many cells thick: the retina (Figure 26).

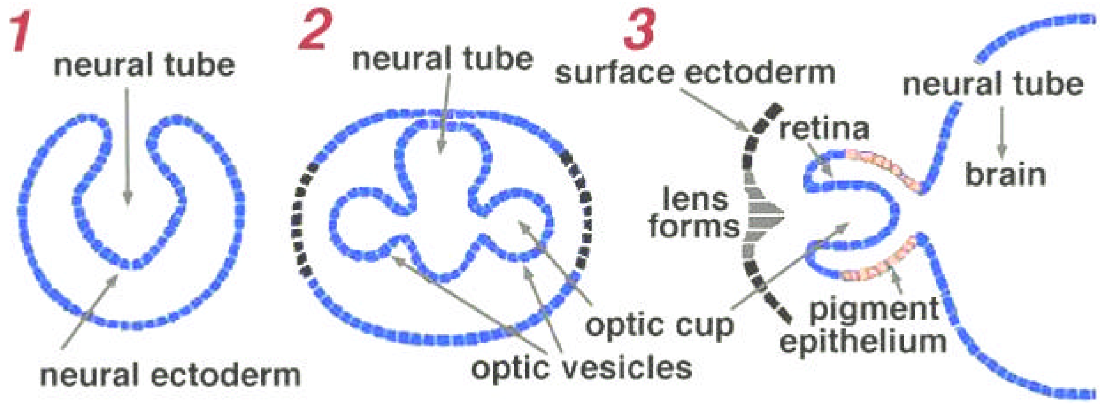


Figure 26. Development of the eye from the neural tube through the optic vesicles and the inverted optic cup forming the retina.
Figure adapted from http://webvision.med.utah.edu/anatomy.html.

Sensory retinal development begins as early as the optic vesicle stage, with the migration of cell nuclei to the inner surface of the sensory retina. Additional retinal development is characterized by the formation of further layers arising from cell division and subsequent cell migration. The retina develops in an inside to outside manner: ganglion cells are formed first and photoreceptors cells become fully mature last. Further changes in retinal morphology are accomplished by simultaneous formation of multiple complex intercellular connections. Thus by 5 months of gestation most of the basic neural connections of the retina have been established.

The functional synapses are made almost exclusively in the two plexiform layers and the perikarya of the nerve cells are distributed in three nuclear layers. Photoreceptor cell maturation begins with the formation of outer segments (OS) containing visual pigment from multiple infolding of the plasma membrane of each cell. Outer segment formation proceeds and the eye become sensitive to light at about 7 months gestation. The final portion of the sensory retina to mature is the fovea, where the ganglion cell layer thickening begins during mid gestation. The outer nuclear layer is also wider here than elsewhere in the retina and consists almost entirely of developing cone cells. The ganglion cell nuclei migrate radially outwards in a circle, leaving the fovea free of ganglion cell nuclei. Cell-cell attachments persist, however foveal cone cells alter their shape to accommodate the movement of ganglion cells. Foveal development continues with cell rearrangements and alteration in cone shape until about 4 years after birth (Hendrickson and Yuodelis, 1984). Surface membranes cover the eye cup and develop into lens, iris and cornea with the three chambers of fluid filled with aqueous and vitreous humors. Surface membranes cover the eye cup and develop into lens, iris and cornea with the three chambers of fluid filled with aqueous and vitreous humors.

6.2 Overview of retinal anatomy and physiology

The retina is approximately 0.5 mm thick and lines the back of eye. The optic nerve contains the ganglion cell axons running to the brain and, additionally, incoming blood vessels that open into the retina to vascularise the retinal layers and neurons (Figure 27). A radial section of a portion of the retina reveals that the ganglion cells (the output neurons of the retina) lie innermost in the retina closest to the lens and front of the eye, and the photoreceptors (the rods and cones) lie outermost in the retina against the pigment epithelium and choroid. Light must, therefore, travel through the thickness of the retina before striking and activating the rods and cones (Figure 27). Subsequently the absorption of photons by the visual pigment of the photoreceptors is translated first into a biochemical message and then to electrical message that can stimulate all the succeeding neurons of the retina. The retinal message concerning the photic input and some preliminary organization of the visual image into several forms of sensation are transmitted to the brain from the spiking discharge pattern of the ganglion cells. All vertebrate retinas are composed of three layers of nerve cell bodies and two layers of synapses. The outer nuclear layer contains cell bodies of the rods and cones, the inner nuclear layer contains cell bodies of the bipolar, horizontal and amacrine cells and the ganglion cell layer contains cell bodies of ganglion cells and displaced amacrine cells. Dividing these nerve cell layers are two neuropils where synaptic contacts occur.

The first area of neuropil is the outer plexiform layer (OPL) where connections between rod and cones, and vertically running bipolar cells and horizontally oriented horizontal cells occur. The second neurophil of the retina, is the inner plexiform layer (IPL), and it functions as a relay station for the vertical-information-carrying nerve cells, the bipolar cells, to connect to ganglion cells. In addition, different varieties of horizontally- and vertically-directed amacrine cells, somehow interact in further networks to influence and integrate the ganglion cell signals. It is at the culmination of all this neural processing in the inner plexiform layer that the message concerning the visual image is transmitted to the brain along the optic nerve.

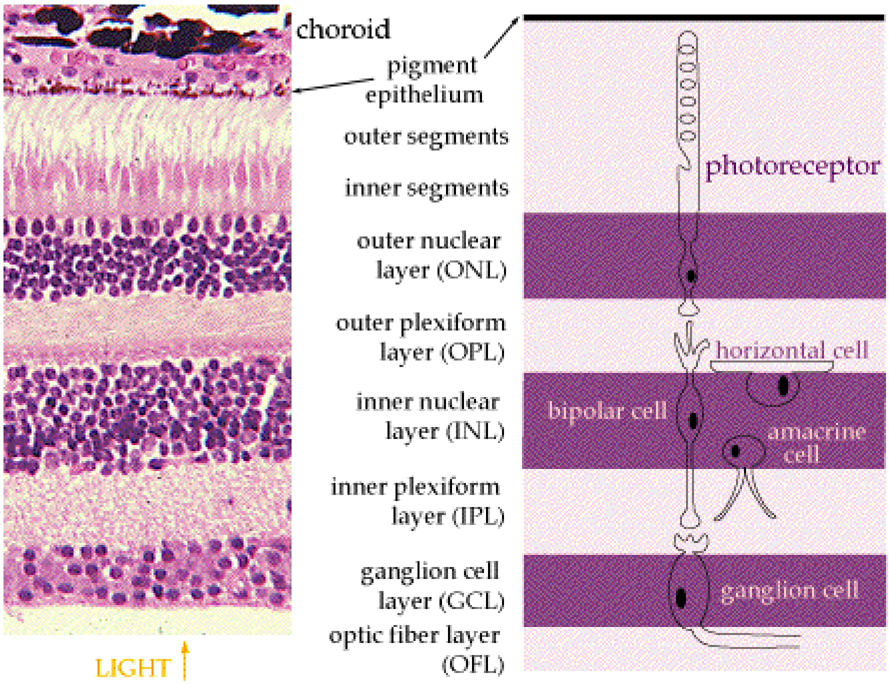


Figure 27. Schema of the layers of the retina. Figure adapted from http://thalamus.wustl.edu.

6.3 Common retinal diseases

The human retina is a delicate organization of neurons, glia and nourishing blood vessels. In some eye diseases, the retina becomes damaged or compromised, and degenerative changes set will eventually lead to serious damage to the nerve cells that carry the vital messages about the visual image to the brain.

One of the great success stories in retinal disease (RD) research in the past decade has been identification of many of the genes and mutations causing inherited retinal degeneration. To date, more than 133 RD genes have been identified (Figure 28), encompassing many disorders such as retinitis pigmentosa, Leber congenital amaurosis, Usher syndrome and macular dystrophy. The most striking outcome of these findings is the exceptional heterogeneity involved: dozens of disease-causing mutations have been detected in most RD genes; mutations in many different genes can cause the same disease; and different mutations in the same gene may cause different diseases.

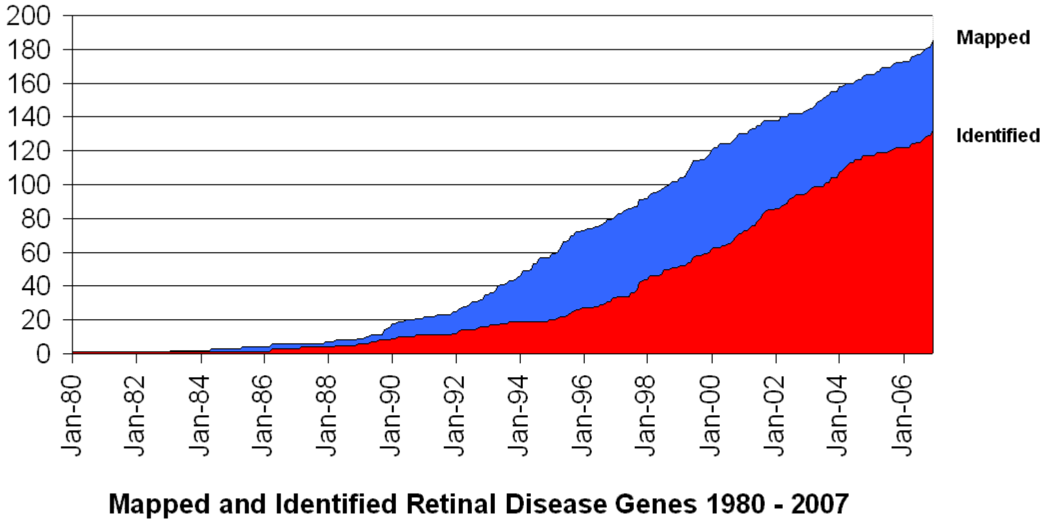


Figure 28. Graphical representation of number retinal disease genes that are mapped (blue) and identified (red). Figure adapted from RetNet http://www.sph.uth.tmc.edu.

Important retinal diseases are described below.

6.3.1 Retinitis pigmentosa (RP)

Retinitis pigmentosa (RP) is a genetically heterogeneous retinal degeneration, characterized by degeneration of rod followed by cone photoreceptors. The first clinical symptoms of RP are night blindness and narrowing of the peripheral field of vision that gradually deteriorate to become “tunnel-like”. Eventually, the central vision is reduced to total blindness in most cases. The worldwide prevalence of retinitis pigmentosa is about 1 in 4000 for a total of more than 1 million affected individuals. The disease can be inherited as an autosomal-dominant (about 30–40% of cases), autosomal-recessive (50–60%), or X-linked (5–15%) trait, the etiology of the disease is still unknown in half of all the patients, illustrating that our understanding of the disease remains limited (Hartong *et al.*, 2006).

6.3.2 Glaucoma

Glaucoma is a heterogeneous group of optic neuropathies that share a similar set of clinical features including characteristic cupping of the optic nerve head and selective apoptotic death of retinal ganglion cells (RGC), which leads to a progressive loss of visual field and blindness. A number of risk factors are associated with the development of glaucoma including age, ethnicity (race), family history (genetics), myopia, and central corneal thickness. However, the most important risk factor for the development and progression of glaucoma is intraocular pressure (IOP) (Pang and Clark, 2007). The front portion of the eye is filled with a nourishing, protective fluid called aqueous humor. This transparent fluid is continuously circulated through the inner eye, flowing in and draining out. If the drainage area of the eye called the Trabecular Meshwork is blocked, the fluid pressure within the inner eye may increase, and resulting is decreased blood supply to the eye's optic nerve and damage. The optic nerve will no longer be able to transfer the information form the retina to the brain. This damage can result in partial or complete blindness.

6.3.3 Age related macular degeneration (AMD)

The central part of the retina is called the macula and is responsible for vision that is needed for reading and other detailed work damage to the macula results in poor vision. The most common disease process that affects the macula is age-related macular degeneration (AMD). AMD is characterized by a progressive loss of central vision attributable to degenerative and neovascular changes in the macula, a highly specialized region of the ocular retina responsible for fine visual acuity. Estimates gathered from the most recent World Health Organization (WHO) global eye disease survey conservatively indicate that 14 million persons are blind or severely visually impaired because of AMD. (Gehrs *et al.*, 2006).

6.3.4 Leber's congenital amaurosis (LCA)

In 1869, Dr Leber described a disorder associated with congenital amaurosis, nystagmus, and the oculodigital sign that appeared to be a variety of retinitis pigmentosa. This disorder, now referred to as Leber's congenital amaurosis (LCA), is a group of autosomal recessive dystrophies with a heterogenous clinical and genetic background (Waardenburg and Schappert-Kimmijser, 1963). To date, mutations of seven genes have been reported to be implicated in the disease: RetGC1 (Perrault *et al.*, 1996), RPE65 (Gu *et al.*, 1997),(Lorenz *et al.*, 2000), CRX (Freund *et al.*, 1998), AIPL1 (Sohocki *et al.*, 2000),(Heegaard *et al.*, 2003), LRAT (Thompson *et al.*, 2001), CRB1 (Lotery *et al.*, 2001), and RPGRIP (Gerber *et al.*, 2001). In addition, two other loci may be involved: LCA3 on 14q24 (Stockton *et al.*, 1998) and LCA5 on 6q11-16 (Dharmaraj *et al.*, 2000).

LCA occurs at an incidence of 3/100,000 newborns and currently no treatment is available. The pathophysiology of LCA is unknown, however, histological data are consistent with abnormal development of photoreceptor cells in the retina and extreme premature degeneration of retinal cells (Heegaard *et al.*, 2003),(Kroll and Kuwabara, 1964; Flanders *et al.*, 1984; Sullivan *et al.*, 1994).

6.3.5 Retinoblastoma (Rb)

Retinoblastoma is a cancer of the retina. Development of this tumor is initiated by mutations (Knudson, 1971) that inactivate both copies of the RB1 gene, which codes for the retinoblastoma protein (Friend *et al.*, 1986). Death is caused by tumour spread to the brain through the optic nerve or long the meninges. The tumour arises in embryonic retinal neuroblasts and can be unilateral or bilateral, growing as a solitary or multifocal tumour. Initially the tumour spreads locally into the vitreous (endophytic growth) or subretinal space (exophytic growth). In either case, the child loses vision and may develop a squint with the appearances of a white mass behind the lens (leukocoria).

It occurs mostly in children younger than 5 years and accounts for about 3% of the cancers occurring in children younger than 15 years. Retinoblastoma arises in 1 in 20, 000 children, it is the most common childhood ocular disease. The estimated annual incidence is approximately 4 per million children. About 40% of cases are inherited, while the others arise from spontaneous somatic mutations.

6.3.6 Diabetic Retinopathy (DR)

The effect of diabetes on the eye is called diabetic retinopathy and it is the most common and most serious eye complication of diabetes, which affects the circulatory system of the retina leading to decreased vision or even blindness. Nearly half of people with known diabetes have some degree of diabetic retinopathy. Patients with diabetes are more likely to develop eye problems such as cataracts and glaucoma, Cataracts develop at an earlier age in people with diabetes and glaucoma is twice more common in diabetics. Most patients develop diabetic changes in the retina after approximately 20 years.

6.4 Models for the congenital retinal disorders

Animal models provide a valuable tool for investigating the genetic basis and the pathophysiology of human diseases, and to evaluate therapeutic treatments. To study congenital retinal disorders, mouse mutants have become the most important model organism. (Dalke and Graw, 2005)

Mice suffering from hereditary eye defects (and in particular from retinal degenerations) have been collected since decades (Keeler, 1924). They allow the study of molecular and histological development of retinal degenerations and to characterize the genetic basis underlying retinal dysfunction and degeneration. The recent progress of genomic approaches has added increasing numbers of such models. One such model is LCA, the Lrat-/- mouse, recapitulates clinical features of the human Leber congenital amaurosis (LCA), Batten and Co-workers (Batten *et al.*, 2005) have successfully used intraocular gene therapy to restore retinal function to Lrat-/- mice. Similarly, mutations in RPE65 lead to a range of retinal dystrophies ranging from from LCA to autosomal recessive retinitis pigmentosa. One of the most frequent missense mutations is an amino acid substitution at position 91 (R91W), concerning this Marijana Samardzija *et al* have generated R91W knock-in mice to understand the molecular mechanisms of retinal degeneration caused by this aberrant Rpe65 variant (Samardzija *et al.*, 2007).

Although mouse models are good tool to investigate retinal disorders, mouse retina are slightly different from human retina, particularly with respect to the number and distribution of the photoreceptor cells. The mouse is a nocturnal animal and has a retina dominated by rods with cones that are small in size and represent only 3-5% of the photoreceptors. Mice do not form cone-rich areas like the human fovea. Instead of three cone pigments present in the human retina, mice express only two distinct pigments with absorption maxima near 350 and 510 nm (Lyubarsky *et al.*, 1999). The other important model is Zebrafish, it fills a gap in the current repertoire of models, offering genetic tractability in a vertebrate. Their retina has many similarities with a human retina. Importantly, unlike rodents, they have rich colour vision, offering the potential to model the macular degenerations (Goldsmith and Harris, 2003). Drosophila is very much at the opposite end of the experimental spectrum to humans. They have a short life cycle, tractable genetics and can be bred in large numbers. But there are fundamental differences between invertebrate and vertebrate phototransduction, limiting their ability to model the full range of human disease. Larger animal models like Chicken (Semple-Rowland *et al.*, 1999), Pig (Petters *et al.*, 1997) and Dog (Acland *et al.*, 2005) for retinal degenerations) also exist. These are particularly useful for establishing efficacy and safety parameters of proposed therapies. It is also much easier to assess potential transplantation and surgical techniques in larger animals.

Gene symbol	**Chr.(cM)**	**Defect alleles**	**Mutation**	**Reference**
Rpe65 retinal pigment epithelium 65	3 (87.6)	*Rpe65*^{tm1Tmr}	Knockout, exons 1–3 of the gene were replaced with a PGK-neo cassette	(Redmond *et al.*, 1998)
		Rpe65^{rd12} (retinal degeneration 12)	Nonsense mutation, base substitution (C to T) in codon 44	(Pang *et al.*, 2005)
Pde6b rod phospodiesterase, beta subunit (r,rodless; rd, retinal degeneration)	5 (57.0)	*Pde6b*^{rd1} (retinal degeneration 1)	Nonsense mutation, C-A transversion in codon 347 (exon 7)	(Pittler and Baehr, 1991)
Rho rhodopsin	6 (51.5)	*Rho*^{tm1Jlem}	Knockout, a PGK-neo cassette was inserted into the first coding exon	(Lem *et al.*, 1999)
		Rho^{tm1Phm}	Knockout, a neomycin cassette under the control of a polymerase II promoter was inserted at codon 135 in exon 2	(Humphries *et al.*, 1997)
Nr2e3 nuclear receptor subfamily 2, group E, member 3 (PNR-photoreceptor-	9 (33.5)	*Nr2e3*^{rd7}	Deletion of 380 bp (exon 4 and 5) - frame shift resulting in a premature stop codon	(Akhmedov *et al.*, 2000)

specific nuclear receptor)				
Nrl neural retina leucine zipper gene	14 (19.5)	*Nrl*^{tm1Asw}	Knockout, a PGK-neomycin resistance cassette replaced the entire coding region (exons 2 and 3)	(Swain *et al.*, 2001)
Rds retinal degeneration, slow (Prph2-peripherin2; rd2)	17 (18.8)	*Rds*$^{Prph2-rd2}$	Insertion of ∼10kb, disrupting the coding sequence in exon 2	(Travis *et al.*, 1989)
RP1h retinitis pigmentosa 1 homolog (human)	1 (6.5)	*Rp1*$^{-/-}$	replaced a 2.5-kb genomic fragment including exons 2 and 3 of the *Rp1* gene with a 1.6-kb DNA fragment containing the neomycin gene.	(Gao *et al.*, 2002)

Table 4. Overview of mutations in the mouse, affecting the structure or function of the retina. For allelic series just a few examples are listed. Transcriptomic experimental data on these mouse models are available in RETINOBASE.

6.5 Retinal transcriptome

The mammalian retina is a valuable model system to study neuronal biology in health and disease. To obtain insight into intrinsic processes of the retina, great efforts are directed towards the identification and characterization of transcripts with functional relevance to this tissue. Comparative analysis of retinal and RPE (Retinal Pigment Epithelium) expression profiles at various stages of development and aging and during the progression of disease pathogenesis is likely to have broad implications for delineating fundamental biological processes and identifying targets for drug discovery.

Over the last decade, several groups have identified genes and expressed sequence tags (ESTs) from the retina and the RPE (Sinha *et al.*, 2000), (Gieser and Swaroop, 1992), (Malone *et al.*, 1999; Bortoluzzi *et al.*, 2000; Stohr *et al.*, 2000), but it is only recently that the retinal transcriptome has been studied with high-throughput molecular and computational tools. Additional useful resources for ESTs that are expressed in the retina at 14.5 days of development is RetinalExpress (http://odin.mdacc.tmc.edu/RetinalExpress). In addition to direct sequencing efforts, several groups have used the serial analysis of gene expression (SAGE) approach (Velculescu *et al.*, 1995) to profile expression in the murine retina (Blackshaw *et al.*, 2001), human retina (Sharon *et al.*, 2002), human cornea (Gottsch *et al.*, 2003), and rat extraocular muscle (Cheng and Porter, 2002). Such analyses have provided valuable new information and insights into the complex processes of retinal development, aging and disease.

Heidi L Schulz et al. defined first reference transcriptome of adult retina/retinal pigment epithelium. They have extracted 13,037 non-redundant annotated genes from nearly 500,000 published datasets on redundant retina/retinal pigment epithelium (RPE) transcripts. The data were generated from 27 independent studies employing a wide range of molecular and biocomputational approaches (Schulz *et al.*, 2004).

Recently, Bowes Rickman C *et al* (Bowes Rickman *et al.*, 2006) developed large-scale, high-throughput annotation of the human macula transcriptome in order to identify and prioritize candidate genes for inherited retinal dystrophies, based on ocular-expression profiles using serial analysis of gene expression (SAGE). Their finding suggested that Cone photoreceptor-associated gene expression was elevated in the macula transcription profiles.

As expression profiles of different tissues and cells and of diseased and/or mutant retinas and RPEs become available, it will be possible to identify transcripts that are present in a single or a limited number of cell types and those associated with disease pathogenesis. Recent analysis of RPE transcriptome by Rizzolo LJ *et al* (Rizzolo *et al.*, 2007) revealed dynamic changes during the development of the outer blood-retina barrier. The data from the above study indicate extensive remodeling of the extracellular matrix, cell surface receptors, cell-cell junctions, transcellular ion transport, and signal transduction pathways throughout development.

A genome-wide expression profiling study of retinoschisin-deficient retina in early postnatal mouse development revealed that microglia/gila activation may be triggering events in the photoreceptor degeneration of retinoschisin-deficient mice. Furthermore, the data point to a role of Erk1/2-Egr1 pathway activation on RS pathogenesis (Gehrig *et al.*, 2007).

Zhang SS et al (Zhang *et al.*, 2006)study provides a complementary genome-wide view of common gene dynamics and a broad molecular classification of mouse retina development. Different genes in the same functional clusters are expressed in different developmental stages, suggesting that cells might change gene expression profiles from differentiation and maturation stages. The large-scale changes in gene regulation during development are necessary for the final maturation and function of the retina.

The microarray studies have ranged from identification of genes altered during aging of human retina, (Yoshida *et al.*, 2002; Chowers *et al.*, 2003b); genes involved in organization of developing retina in embryonic mice (Diaz et al., 2003) identification of a disease-causing gene based on differential expression between wild-type and rhodopsin knockout mice (Kennan *et al.*, 2002) and identification of novel genes that may be preferentially expressed in the retina (Chowers *et al.*, 2003a).

Genome-wide expression profiling with microarrays is expected to yield tremendous amounts of data. Efficient mining of these data is currently a challenge for vision scientists and computational biologists. It is critical to develop acceptable standards for microarray experimentation and guidelines for data presentation, storage, and sharing. In this regard, we had developed RETINOBASE: the Retina Gene Expression Database which is a new unifying resource for retinal gene expression data, specifically developed to respond to the needs of researchers working in the field of retinal biology, which is described in detail in the chapter 10. In addition, although all the studies have made significant contributions to ophthalmic research, greater success could be achieved if specialized arrays that contain all the genes expressed in the eye are used. With reference to this, taking advantage of RETINOBASE, we have developed RETCHIP (an oligonucleotide microarray composed of 1500 genes relevant to retinal biology) to speed and systematize the study of mouse retinal development and degeneration.

Materials and Methods

Chapter 7: Informatic and Bioinformatic resources

The developments and analyses described in chapters 7-10 were performed used the existing infrastructure and computer resources of the BioInformatics Platform of Strasbourg (BIPS) (http://bips.u-strasbg.fr) of the IGBMC. It is equipped with important and high quality facilities for data storage and calculations. The BIPS is a high-throughput platform for comparative and structural genomics, which was identified in 2003 as a national inter-organizational technology platform (*Plate-forme Nationale RIO).* The Bioinformatics Platform is part of the "Strasbourg Alsace-Lorraine-Génopole®" and the "Cancéropôle du Grand-Est". It was labelled "Plate-forme Nationale RIO" in 2003 and in 2006 it has been listed as a "European Research Infrastructure" by the "European Science Foundation".

7.1 Informatics resources

7.1.1 Calculation and data storage options

The central servers are currently available for database development and computational data analyses:

(i) Interactive and web services: Sun Enterprise 450 (Solaris 9). Four processors with 1 Gb shared memory.

(ii) Computational servers:

- ❑ Six Compaq ES40 cluster (Tru64 UNIX). 6 x 4 EV67 processors. Of the six machines in this cluster, five have 4 Gb memory each, and the sixth has 16 Gb.
- ❑ Eight Sun V40z servers (2 x Solaris 10 and 6 x RedHat Enterprise Linux 4). 6 x 4 Opteron processors with 2 x 32 Gb and 4 x 16 Gb memory.

(iii) Disk server: Sun V480 (Solaris 9) providing 8 Terabytes on Raid5 disks shared with other servers using NFS.

(iv) Linux Ubuntu on 64 bit AMD Opteron on biprocessors or quadriprocessors.

We also used a PC with Windows XP to run the native ODBC (Open DataBase Connectivity) drivers for Excel or Access. We maintain a nearby cloned system for security, backup, restoration and test reasons.

7.2 Data sources: Gene Expression Omnibus database

There are several ways to retrieve GEO data. One way is by entering a valid GEO accession number in the Accession Display bar or by querying GEO "DataSets" by entering a valid search term for example "Retina" the query returns all the datasets concerning the retina. Another way is to browse the list of current GEO repository contents. All data are available for download from the GEO FTP site.

All public datasets in the form of CEL files or signal intensities are present in RETINOBASE and almost all the datasets used in the meta-analysis by MDF were downloaded via FTP from Gene Expression Omnibus (GEO) (Barrett *et al.*, 2006).

7.3 Programming languages

The RETINOBASE database website is written in PHP5 (Hypertext Preprocessor), dynamically creating HTML (HyperText Markup Language) pages with JavaScript and CSS (Cascading Style Sheets).

Our implementation is mainly based on open source or free programs and languages. Thus for core RETINOBASE and website, no specific requirements concerning computer architectures or systems are needed as long as Apache, the DBMS (Database Management System) PostgreSQL and PHP language are running on the chosen platform.

We used PHP because of its following advantages

- ❑ PHP provides easy environment to work. PHP also provides object oriented (OO) features that enable to design modern Web-based applications that are robust and secure.
- ❑ PHP runs virtually on any operating system.
- ❑ PHP natively integrates with a large array of database engines, both open-source (MySQL, PostgreSQL, SQLite) and commercial (Oracle, MS SQL Server).
- ❑ PHP is free open source software. There are a number of websites, blogs, discussion forums, and other resources available to PHP developers. There are also thousands of readily available PHP-based applications, libraries of functions, components, and frameworks that can be used to develop applications more quickly.
- ❑ We have in-house expertise in PHP.

We have used Tcl/TK to retrieve all the orthologs Probe Sets (in different GeneChip arrays) corresponding to a given Probe set in a given GeneChip array from orthologs the file are available at www.affymetrix.com

7.3.1 Normalization software

Public and propriety data obtained at the level of CEL files have been analysed with three different normalization programs – RMA (Irizarry *et al.*, 2003b), dChip (Li and Wong, 2001) and MAS5 (Hubbell *et al.*, 2002). RMA (Robust Multichip Average) is a probe level analysis algorithm for Affymetrix Gene Chip. It consists of three steps: a background adjustment, quantile normalization (Bolstad *et al.*, 2003) and finally summarization. RMA uses a stochastic model to estimate gene expression. This algorithm uses the probe data stored in Affymetrix CEL files as input and converts the probe level expression data into gene level expression data. RMA preprocesses and normalizes the data to a certain extent in such a way that the distribution of expression values is comparable across the different samples. dChip (DNA-Chip Analyzer) is a Windows software package for probe-level and high-level analysis of Affymetrix gene expression microarrays and SNP microarrays. At the probe level, dChip can display and normalize the CEL files, and the model-based approach allows pooling information across multiple arrays and automatic probe selection to handle cross-hybridization and image contamination. Using R statistical package (http://www.r-project.org) and Bioconductor (Gentleman *et al.*, 2004). The signal intensities thus obtained were integrated into RETINOBASE.

7.3.2 R and Bioconductor

R is a language and environment for statistical computing and graphics (http://www.r-project.org). It was originally created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is now developed by the R Development Core Team. It is a GNU project, which is similar to the S language and environment that was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data. It is primarily based on the R programming language. Although initial efforts focused primarily on DNA microarray data analysis, many of the software tools are general and can be used broadly for the analysis of genomic data, such as SAGE, sequence, or SNP data.(Gentleman *et al.*, 2004).

The broad goals of the projects are to:

- ❑ Provide access to a wide range of powerful statistical and graphical methods for the analysis of genomic data.
- ❑ Facilitate the integration of biological meta-data in the analysis of experimental data: e.g. literature data from PubMed, annotation data from LocusLink.
- ❑ Allow the rapid development of extensible, scalable, and interoperable software.
- ❑ Promote high-quality documentation and reproducible research.
- ❑ Provide training in computational and statistical methods for the analysis of genomic data.

Further, we have generated quality control reports using affyQCReport- an R package that generates quality control reports for Affymetrix array data (Gautier *et al.*, 2004; Wilson and Miller, 2005) for all experiments where .CEL files are available.

7.4 DAVID software for gene ontology analysis

We used the DAVID software (Dennis *et al.*, 2003) (described in chapter 3) to obtain gene ontology enrichment for our protein and gene lists in chapter 8 and 9 respectively.

DAVID accepts a wide range of gene accessions or IDs (Affymetrix probeset id, gene name, symbol, RefSeq transcript and protein id, gene bank accession, entrez gene id etc.) to query the database. Our protein/gene lists were uploaded in plain text format and the annotated lists from DAVID were downloaded in tab-delimited format.

Results and Discussion

This chapter will present the work and the results we have obtained in three different projects which encompass two major aspects of high throughput data analysis namely, the development of algorithms and the creation and implementation of a new relational database in this context RETINOBASE.

As our laboratory is involved in the development of various algorithms for improved clustering methods, meta-analysis and statistical analysis, I had the opportunity to participate in the last phase of algorithmic development, corresponding to the application of a newly developed algorithm to biological data and analysis of the results. In this context, the first project concerned the development of a Maximum Likelihood Approximation (MLA) method for Dirichlet's parameter estimation. In this project, I was mainly involved in the application of the MLA method to protein sequence data and interpretation of the results, taking advantage of the Gene Ontology analysis tools and strategies implemented in RETINOBASE. After a rapid description of the method, chapter 8 will present some of the main results obtained that will be further discussed. In the second project, I was mainly involved in the application of a newly developed method called MDF for Multi-Dimensional Fitting on transcriptomic datasets and in the analysis and interpretation of the results. This is detailed in chapter 9.

The final project concerns RETINOBASE, the web database, data mining and analysis platform for gene expression data on retina. RETINOBASE is a new relational database, implying that I had to tackle all aspects of database creation, including database design and implementation, data mining for retinal specific experiments in public data repositories such as GEO and ArrayExpress, transcriptomic data treatment and clustering analysis and also, creation of a user-friendly interface. The utilities and some of the functionalities are described in detail in chapter 10.

Chapter 8. A Maximum Likelihood Approximation method for Dirichlet's parameter estimation (Publication 1)

8.1 Scientific context

As the information embedded in high-throughput experiments appears to be entangled in a complex mix of various types of noise, a variety of statistical methods to reduce this noise and aid in data analyses is needed in order to extract valuable information.

In general, Dirichlet distributions are natural choices to analyse data described by frequencies or proportions since they are the simplest known distributions for such data, apart from the uniform distribution. Parameter estimation is usually performed with the Newton-Raphson algorithm after an initialisation step using either the moments or the Ronning's methods. However this initialisation can result in parameters that lie outside the admissible region. In order to avoid this, and to accelerate the convergence of optimization algorithms to the optimum, a novel, simple and very efficient method for estimating the parameters of a Dirichlet distribution by maximum likelihood was developed by Nicolas Wicker (described in publication number 1).

My contribution to the MLA development was mainly linked to the analysis of the advantages of the MLA method over the moments methods using a biological example. We decided to compare the different methods for the clustering of protein sequences based on their amino acid compositions in a mixture model framework. By means of Gene Ontology (GO) annotations, we further went on to the interpretation of the biological relevance and we identified GO term enrichments in the clusters obtained. We took advantage of the hypergeometric distribution (Spellman and Rubin, 2002), (Jensen *et al.*, 2003) a commonly used enrichment method to determine enriched GO (Gene Ontology) terms in a given data.

Even though amino acid composition is highly variable among the species, amino acid composition has already been shown to be biologically meaningful and particularly relevant in identifying function as well as localization (Cedano *et al.*, 1997). Notably, some specific protein families exhibit very biased and conserved amino acid composition that can be useful to provide functional guidelines in cross species identification of families of unknown protein.

In this chapter, I will mainly focus on the clusters that are found using the Dirichlet distribution.

8.2 Datasets used and identification of clusters

We have extracted 13 432 well annotated human protein sequences from Swiss-Prot (Boeckmann et al., 2003) and applied the mixture model method with Dirichlet distributions. As mentioned earlier, these datasets were clustered in a mixture model framework with the MLA initialisation and alternatively with moments method. In both cases the number of clusters was determined using AIC criterion (Akaike, 1974). We then estimated the optimum number of clusters obtained for MLA and moments methods to be 95 and 60 respectively; we noticed that this difference is mainly linked to the fact that moments method frequently failed owing to minimal variation in one or more clusters.

Among the different cluster variations, we identified 4 specific clusters (Figure 29) and (Figure 30) that are unique to the MLA method and not to the moments method. The first cluster, rich in proline (P), has its phenylalanine (F) and tryptophan (W) variance equal to 0. The second and third clusters are rich in serine and arginine (S and R), and in lysine (K) respectively and small variances are observed for aspartate (D) and tryptophan (W). The fourth cluster (not presented in publication 1) is rich in cysteine and serine (C and S) while small variances are observed for numerous residues such as the aliphatic and aromatic ones (I, L and V) as well as for some charged residues (D, N and H).

The biological relevance of these clusters has been assessed using the GO enrichment and KEGG pathway analysis. This was done using the GO (Gene Ontology) annotations through DAVID (the Database for Annotation, Visualization and Integration Discovery) and AmiGO (Ashburner *et al.*, 2000) and a hypergeometric model was used to associate a *p*-value with each term (Cho *et al.*, 2001); the smaller the *p*-value, the more significant are the over-representations and, usually, a *p*-value lower than 0.05 is considered as significant.

8.2.1 Biological significance of the clusters

The **first cluster** (cluster size=9) shown in (Figure 29) is particularly interesting because it is unidentifiable using the moments method as the variance is equal to zero for the phenylalanine and tryptophan (F and W, respectively).

Additionally, the cluster is biologically relevant and significant, with a *p*-value equal to 0.002 for the GO annotation "wound healing" (GO:0042060). The cluster includes nine small proline-rich proteins (SPRR), a novel class of polypeptides that are strongly induced during differentiation of human epidermal keratinocytes *in vitro* and *in vivo*. SPRRs not only provide resistance and flexibility to specialised tissues but also play a key role in the adaptation of epithelial barriers to a large variety of endogenous and external stimuli (Cabral *et al.*, 2001).

We observed that a larger cluster (cluster size=23) obtained with the moments method contains 8 SPRR proteins, along with other proline rich proteins such as salivary and basic proline-rich proteins as well as basic proline-rich peptides.

In **cluster 2** (cluster size=29), we observed that GO annotation "RNA splicing" (GO:0008380) and "nuclear mRNA splicing, via spliceosome" (GO:0000398) are enriched with a *p*-value equal to 0.002 and 0.00000248, respectively (Figure 29). This cluster shows a strong presence of proteins involved in splicing mechanisms (18 out of 29), such as spliceosome formation and mRNA splicing, and more specifically the family of splicing factors that are serine/arginine rich (SR). SR proteins belong to a larger family of polypeptides with "alternating arginine" domains. These splicing factors play an important role in constitutive and regulated pre-mRNA splicing, acting as driving forces during spliceosome assembly. In addition, they play a crucial role in alternative splice-site selection, signifying that they are critical players in the regulation of splicing during cell differentiation and development (Valcarcel and Green, 1996). The KEGG pathway identified in this cluster is "Spliceosomal Assembly" (*p*-value=0.019). The moments method found only a few splicing factors (5) in a single cluster of size 50.

Cluster 3 (Figure 29) is enriched with "chromosome organisation and biogenesis" GO annotation (GO:0051276, *p*-value = 0.002). This cluster mainly contains histones (25 out of 39) and 6 ribosomal proteins corresponding to the 60S ribosomal proteins L14, L23A, L24, L29, the 40S ribosomal protein S25 and the Signal recognition particle 14 kDa protein (ribonucleoprotein related). In-depth analysis of the profile presented in (Figure 29), reveals that, in addition to the high content in lysine residues, this cluster also exhibits high content in small residues such as proline, alanine, glycine and serine. This highly biased composition, involving enrichment of lysine and small residues, is a well known signature of the histone family. Histones are responsible for the structure of chromatin and play important roles in the regulation of gene expression. One of the important roles of histones is to package and compact the DNA. The other group of proteins in this cluster is constituted of ribosomal proteins, which are involved in the cellular process of translation. All these protein families are directly involved in strong interaction with nucleic acids which are highly negatively charged molecules thus explaining the over-representation of basic residues observed. In contrast, in the corresponding cluster of size 63 obtained with the moments method, there are only some of the histones (12) and one ribosomal protein. The KEGG pathway identified in this cluster is “Ribosome Assembly” with a *p*-value equal to 0.00012.

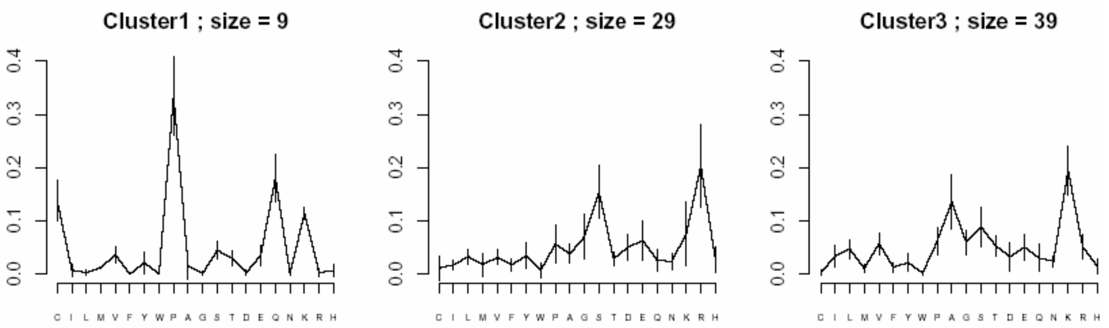


Figure 29. Three clusters found using the MLA method and not found with the moments method.

Along the X-axis are the 20 amino acids denoted by their one letter code: C (cysteine), I (isoleucine), L (leucine), M (methionine), V (valine), F (phenylalanine), Y (tyrosine), W (tryptophan), P (proline), A (alanine), G (glycine), S (serine), T (threonine), D (aspartic acid), E (glutamic acid), Q (glutamine), N (aspargine), K (lysine), R (arginine) and H (histidine). Y-axis indicates the level of variance of each amino acid.

Apart from the above clusters, we also identified an additional cluster (Figure 30), not presented in the publication, and specifically identified by the MLA method; cluster 4 (cluster size=43), which is rich in cysteine and serine (C and S respectively). This cluster mainly consists of keratin associated proteins (23 out of 43) and metallothionein proteins (9 out of 43). GO analysis revealed the molecular function “metal ion binding” (*p*-value=0.00017), which is due to the presence of metallothionein proteins and the cellular component, "keratin filament" (*p*-value=7.1E-44) due to the presence of keratin associated proteins. The corresponding cluster in moments method has a similar number of keratin associated proteins, but does not contain any metallothionein proteins.

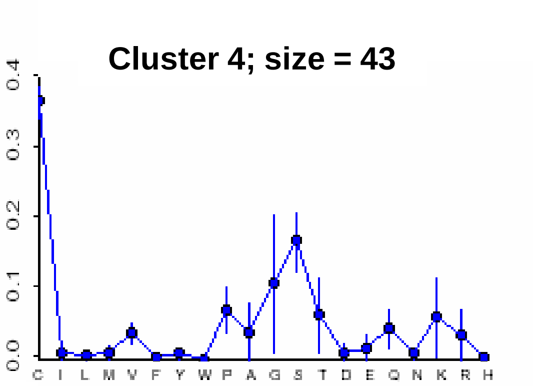


Figure 30. Additional cluster obtained by only MLA method.
Along the X-axis are the 20 amino acids denoted by the one letter code (see Figure 29 legend). Y-axis indicates the level of variance of each amino acid.

Together, these results show the advantages of using the MLA method rather than the moments method when estimating the Dirichlet distribution parameters. The MLA method makes it possible to isolate clusters with a proportion that does not vary. As a consequence, MLA can find many more profiles than the moments method, which stops when one cluster has a small variance.

8.3 Discussion

We further plan to apply this method in MACSIMS (multiple alignment of complete sequences information management system). We think this method will aid in a better partitioning of subfamilies which is based on the multiple alignment of complete sequences, which in turn aids in propagating the known information (Pfam domain, active site residues) to all the sequences in the alignment depending upon conservation criteria.

Another application would be to apply Dirichlet mixture models on transcriptomic data. However, to take advantage of the MLA method, since the Dirichlet distribution is suitable only for proportion data; gene expression levels need to be standardized (described in annexe 1) to proportion data. We do not know yet the significance of the results, since the above standardization has to be fully tested. This standardization should of course be compared with the one where the mean is set to 0 and variance to 1.

Publication n^{o} 1

ELSEVIER
Volume 52, issue 3
1 January 2008
ISSN 0167-9473
COMPUTATIONAL STATISTICS & DATA ANALYSIS
Incorporating Statistical Software Newsletter
The official journal of
iasc
The International Association for Statistical Computing
A Section of The International Statistical Institute
Available online at
ScienceDirect
www.sciencedirect.com

A maximum likelihood approximation method for Dirichlet's parameter estimation

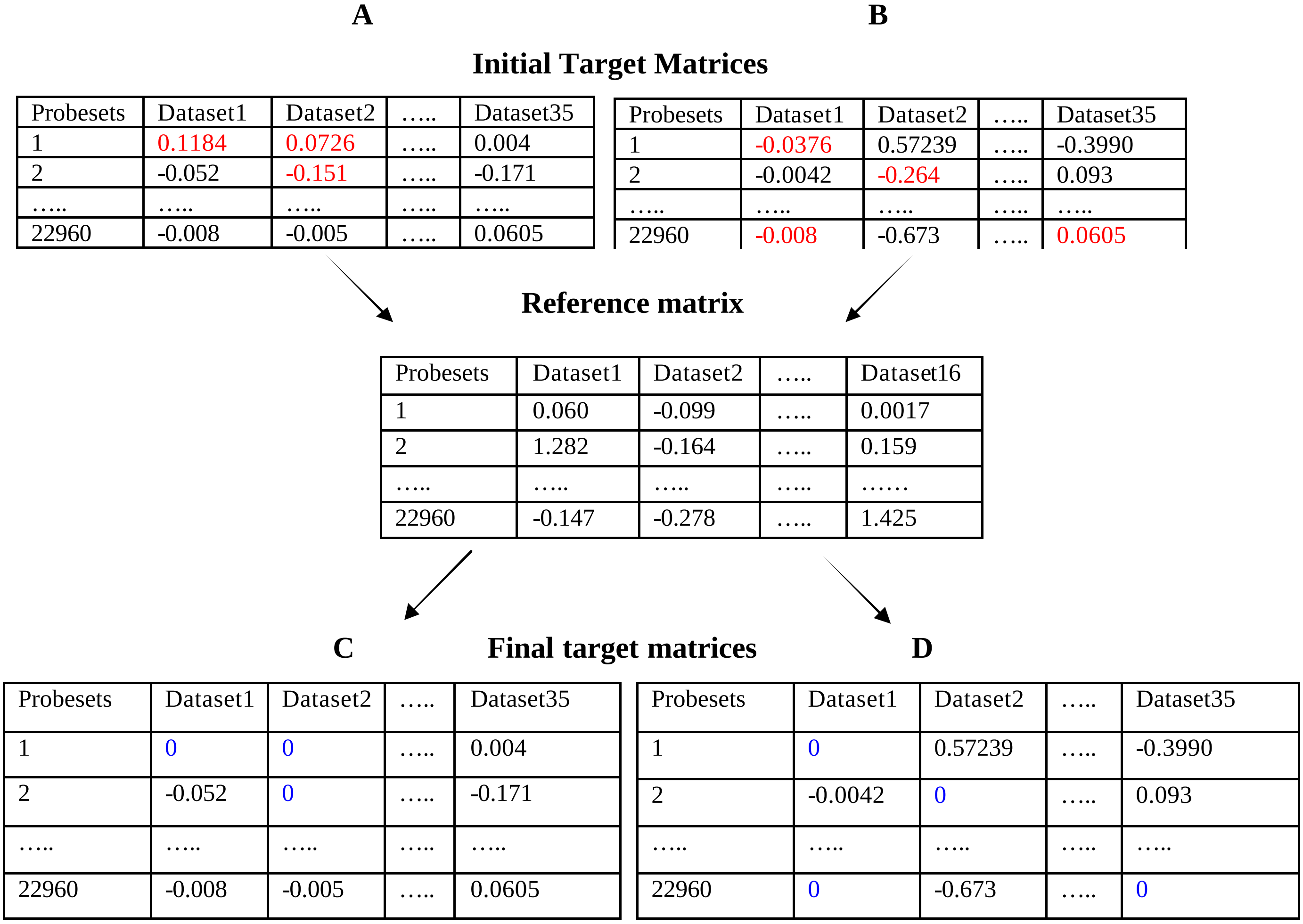
Nicolas WICKER, Jean MULLER, Ravi Kiran Reddy KALATHUR and Olivier POCH

Computational statistics & data analysis, 2008, Vol. 52, N° 3, Pages 1315-1322

Chapter 9. Multi-Dimensional Fitting for transcriptomic data analysis (Publication 2)

9.1 Scientific context

Meta analysis includes a set of statistical techniques for combining information from different studies to derive an overall estimate. In modern biology, it is frequent to deal with large multi-dimensional data matrices. However, statistical methods often have difficulties to efficiently deal with such matrices, because of the noise they inherit. Variable selection and dimension reduction methods are often used to reduce the matrices complexity but this is done at the expense of information conservation. In this chapter, we present the results obtained with a new class of methods called MDF (Multi-Dimensional Fitting) used in the special case where two matrices (with different coordinates) are available to describe the same population (more details about the method and synthetic example are presented in publication number 2). This new class of methods transforms one matrix, designated as target matrix, to fit the distances computed on it with the distances of the second matrix, called reference matrix. In MDF treatment, the "transformation" of a target matrix is obtained by a set of shifts of the initial target matrix values to obtain closer distances with the values observed in the reference matrix depending upon the comparison of all the distances computed on the target and reference matrices (Figure 31). Depending on the type of data handled; discrete or continuous transformation can be performed, in a discrete transformation values are either shifted to 0 or retained as they are initially, while in continuous transformation the values are shifted in a range from 0 to 1. Taking advantage of the results obtained with synthetic data, in our application of the MDF to biological data (see below), we chose the discrete type of transformation because it is easier to partition large datasets and take less computation time in comparison to continuous transformation.



A

Probesets	Dataset1	Dataset2	…..	Dataset35
1	0.1184	0.0726	…..	0.004
2	-0.052	-0.151	…..	-0.171
…..	…..	…..	…..	…..
22960	-0.008	-0.005	…..	0.0605

B

Probesets	Dataset1	Dataset2	…..	Dataset35
1	-0.0376	0.57239	…..	-0.3990
2	-0.0042	-0.264	…..	0.093
…..	…..	…..	…..	…..
22960	-0.008	-0.673	…..	0.0605

Reference matrix

Probesets	Dataset1	Dataset2	…..	Dataset16
1	0.060	-0.099	…..	0.0017
2	1.282	-0.164	…..	0.159
…..	…..	…..	…..	……
22960	-0.147	-0.278	…..	1.425

C

Probesets	Dataset1	Dataset2	…..	Dataset35
1	0	0	…..	0.004
2	-0.052	0	…..	-0.171
…..	…..	…..	…..	…..
22960	-0.008	-0.005	…..	0.0605

D

Probesets	Dataset1	Dataset2	…..	Dataset35
1	0	0.57239	…..	-0.3990
2	-0.0042	0	…..	0.093
…..	…..	…..	…..	…..
22960	0	-0.673	…..	0

Figure 31. Illustration of the general principle of the MDF in the context of transcriptomics target and reference matrices. Values (in red) in initial target matrices (A and B) are optimized to the values in final target matrices (C and D) using MDF based on the values in the reference matrix and shifted to 0 (in blue).

9.2 Datasets

In order to test the new MDF method in a biological application, we have chosen transcriptomics data which induce very large and complex matrices particularly suited for understanding the strengthes and weaknesses of the MDF method. We took advantage of the numerous transcriptomics data available in RETINOBASE and concerning the retina, which is strongly related to neuronal tissues. Thus, we decided to test MDF on neuronal transcriptomics data composed of mixed samples (hereafter called datasets) from retinal and brain origin (mined from RETINOBASE or from publicly available gene expression repositories) versus non-neuronal data, mainly composed of muscle, testis and lung tissues (see Table 2, page 5 in publication 2) (annex 2). We have constructed two target matrices with 35 microarray datasets each, from neuronal and non-neuronal tissues and the reference matrix was composed of 16 datasets from an experiment studying the effects of spontaneous sleep and prolonged wakefulness on the mouse brain (GEO: GSE6514) (Mackiewicz *et al.*, 2007). Though this experiment was performed on Affymetrix Mouse Genome 430 2.0 comprising around 45,000 probesets, we only took probesets that are common to Affymetrix Mouse Genome 430A 2.0 arrays in order to have same number of probesets on both reference and target matrices. All the non-retinal datasets used in this study (presented in detail in publication 2), are publicly available in Gene Expression Omnibus (GEO) (Barrett *et al.*, 2007) and were generated using Affymetrix Mouse Genome 430A 2.0 arrays (http://affymetrix.com) that contains a total of 22,960 probesets. In order to pick datasets for the analyses, we used affyQCReport (Wilson and Miller, 2005) (a package to generate quality control reports for the Affymetrix array data) and picked those datasets which showed highest correlation in correlation heat map.

The reference matrix is of major importance in the context of the MDF treatment thus implying that major care must be taken. We chose the experiment dedicated to “Macromolecule biosynthesis: a key function of sleep” (Mackiewicz *et al.*, 2007) where they studied temporal changes in gene expression during spontaneous sleep and sleep deprivation in the mouse cerebral cortex and in hypothalamus. In this study, they have found remarkable changes in the steady state level of various transcripts during sleep, about ~2,000 genes changed their expression during sleep and found enrichment in some general Gene Ontology terms such as biosynthesis, macromolecular metabolism and also some specialized biological process such as RNA splicing, cholesterol metabolism, etc. We selected this particular experiment for various reasons: firstly, this is a single experiment with a sufficiently large number of datasets (16), secondly, each dataset has 5 replicates, and thirdly, this study addresses possible functions of sleep in a more general way and is not dedicated to any specific problem.

Our hypothesis was that, after transformation of the target matrices through MDF, the probesets exhibiting few shifts to 0 out of the 35 datasets will help us to pinpoint common molecular events occurring in target and reference matrices. The probesets of the target matrix that behave very differently in the reference matrix may require many shifts to fit with the reference matrix data. Thus, we aim to investigate which of the genes from a large heterogeneous collection of expression data from the target matrices from neuronal or non-neuronal origins contain information related to the reference matrix represented by data from brain origin in a sleep deprivation experiment.

9.2.1 Pre-processing of the datasets

All the data sets are obtained at the level of CEL files so that we can perform our own pre-processing steps. In order to reduce or bring the background noise to the same level in both the target matrices, we pre-processed the data using standard treatments. First, we applied the Robust Multi-array Average (RMA) algorithm (Bolstad *et al.*, 2003). We then calculated the ratio (using corresponding control samples as denominator), then the ratios are log2 transformed and normalized using quantile-quantile method. All procedures were performed in R using the packages from Bioconductor (Gentleman *et al.*, 2004). MDF was applied on both the target matrices (neuronal and non-neuronal) separately using the same brain data as the reference matrix.

9.3 Analysis of the results obtained after MDF transformation

As illustrated in Figure 31, after passing through discrete MDF, the initial values of a dataset are either shifted to 0 or retained as they are initially. In both the transformed matrices, the number of probesets that shifted to 0 stands approximately at 64%. We then analysed the number of probesets shifted to 0 per individual dataset in both the neuronal and non-neuronal target matrices and we found that the distribution was quite homogenous ranging from 13018 to 17687 (Figure 32), indicating the homogeneity between the experiments of each target matrix though the individual experiments performed under different conditions.

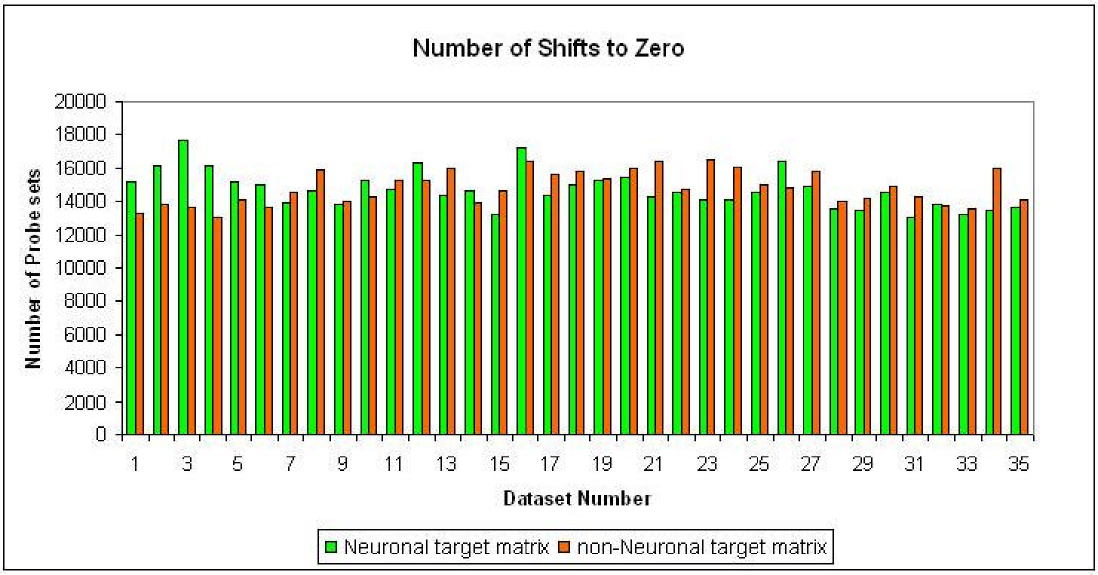


Figure 32. Number of probesets that have at least one shift to zero after MDF. X-axis indicates the dataset number and Y-axis indicates the number of shifted probesets.

Identification of probesets that showed minimum number of shifts to zero

Since the MDF algorithm shifts values that are considered non-informative to 0, we focus our analysis on probesets with very few transformations. When we plotted the cumulative number of shifts against number of neuronal, non-neuronal and common probesets, the common cumulative curve (represented in blue in (Figure 33)) indicates that there are less probesets that are common to both target matrices at lower shifts. Therefore, we initially decided to consider probesets with (0-15) shifts. The number of probesets that showed 0-15 shifts in neuronal and non-neuronal target matrices are 2642, 2914 respectively and the number of common probesets is 410.

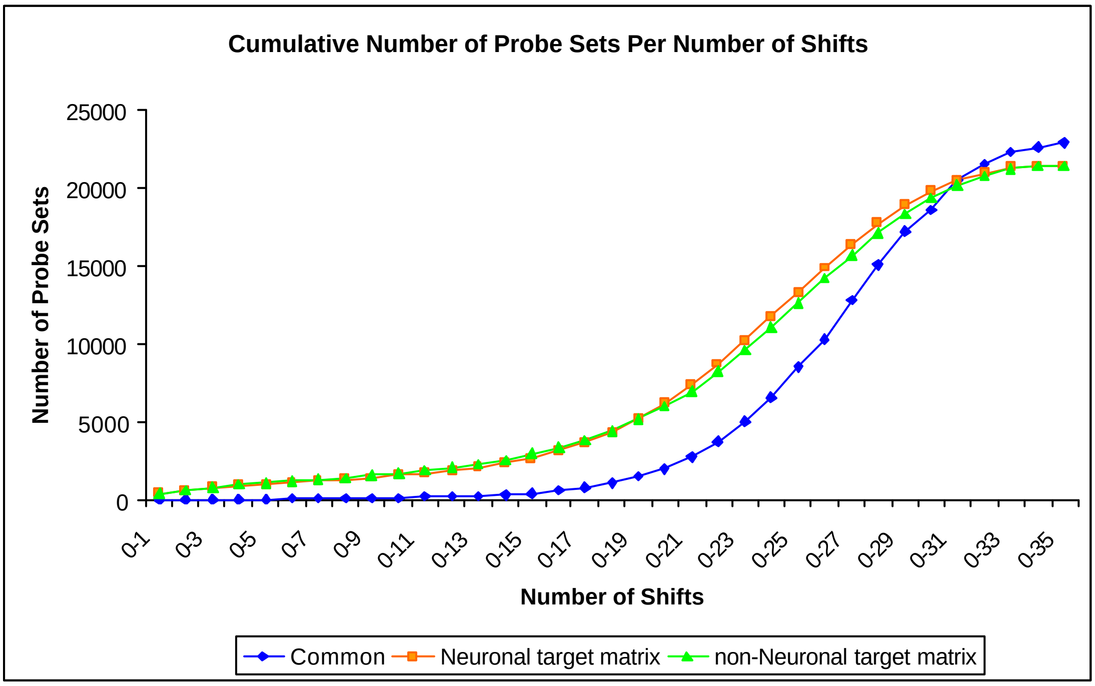


Figure 33. The cumulative curves of the number of probesets respective to the number of shifts are plotted for the neuronal (orange) and the non-neuronal (green) target matrices and for the probesets that are common to both target matrices (blue).

X-axis indicates number of shifts in cumulative distribution and Y-axis indicates number of probesets.

Further in order to have a pertinent cut off value of the number of shifts distinguishing the two target matrix datasets, we went on to plot the frequency distribution of the number of probesets versus the number of shifts in an individual frequency manner rather than cumulative for both target matrices as histograms. We found a bimodal distribution with a main mode centered around 26 shifts covering approximately 85% of all probesets and a small mode on 0 to 6 shifts (Figure 34). The group of probesets that underwent 0-6 shifts in the target matrices was thus selected for further analysis.

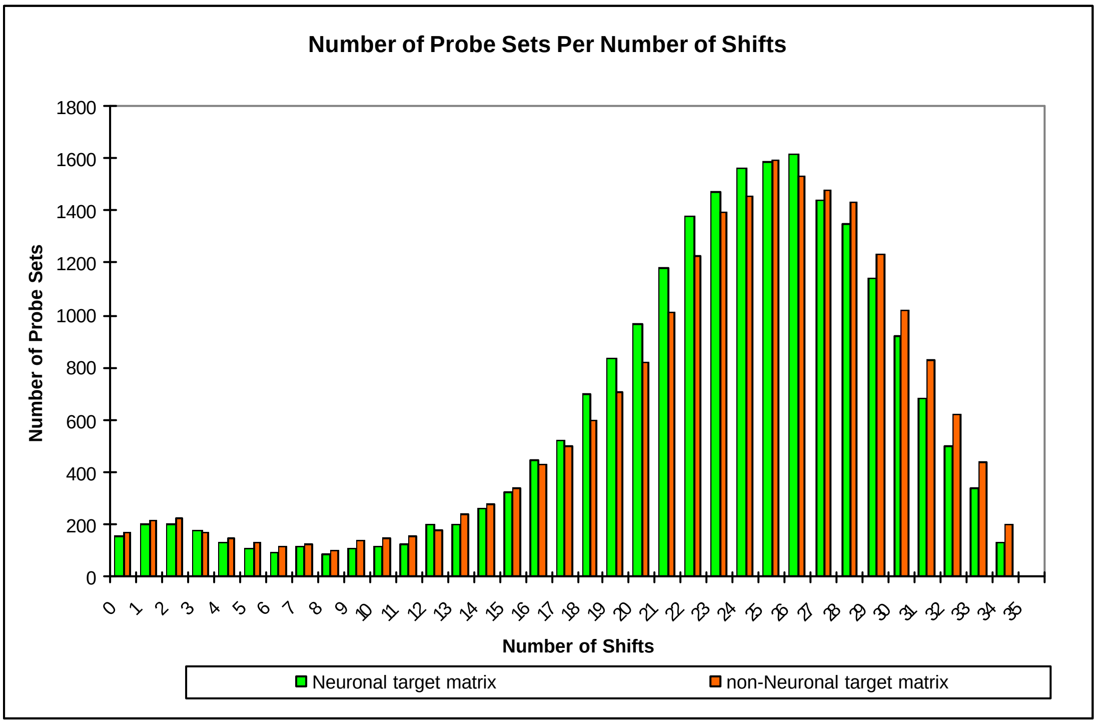


Figure 34. After MDF, the number of probesets per number of shifts is plotted as a histogram for the neuronal and the non-neuronal target matrices.
X-axis indicates number of shifts and Y-axis indicates number of probesets**.** Histogram indicates bimodal distribution of data with one mode around 26 shifts and the other at 0-6 shift. Probesets in group of 0-6 shifts are taken for further functional analysis.

9.4 Bio-analysis of the probesets exhibiting 0 to 6 shifts

Table 5 illustrates the number of probesets present in each cumulative frequency from (0-1) to (0-6) shifts. From the above observation, we picked probesets that fall into the 0-6 shifts category which number 1090, 1185 for neuronal and non-neuronal target matrices and 63 common probesets (Figure 35).

Cumulative number of shifts	Number of probesets (neuronal target matrix)	Number of probesets (non-neuronal target matrix)	Number of probesets common
0-1	365	384	5
0-2	572	613	13
0-3	750	784	25
0-4	883	936	39
0-5	996	1070	48
0-6	1090	1185	63

Table 5. Indicates number of probesets per cumulative shift in both neuronal and non-neuronal target matrices and number of probesets in common.

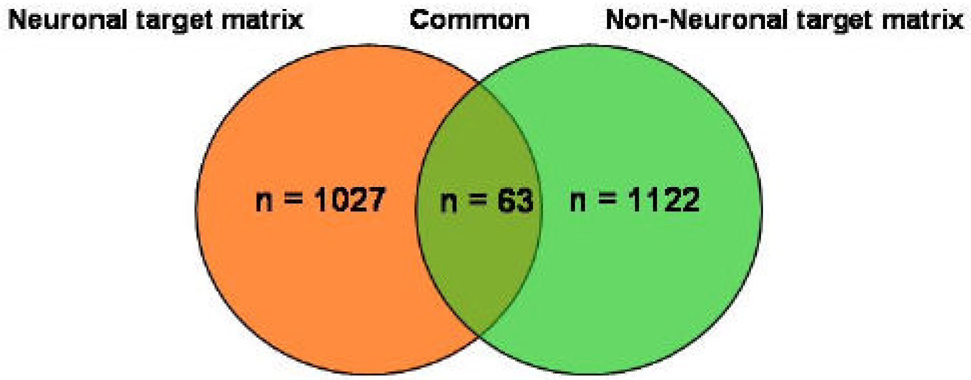


Figure 35. Venn diagram representing the number of probesets present in 0-6 shifts group in both the target matrices and common probesets.
n= number of probesets in 0-6 shifts group.

9.4.1 Functional analysis of the 0-6 shifts group

We went on to investigate GO enrichments using DAVID (Dennis *et al.*, 2003) and AmiGO (Ashburner *et al.*, 2000) to obtain information such as biological process and molecular function concerning the genes present in the 0-6 shifts group (Figure 35). Although many genes in the 0-6 shifts group could be assigned to at least one ontology term, a significant fraction i.e., 37% of 1027 probesets in the neuronal target matrix, 35% of 1122 probesets in the non-neuronal target matrix and half of the common probesets are unclassified, as their function(s) remain to be determined.

Filter criteria for GO annotations

After analysing the remaining 647 probesets for the neuronal target matrix by DAVID, we obtained around 181 GO terms. In the case of the non-neuronal target matrix the remaining 783 probesets correspond to 150 GO terms. For the 36 common probesets we obtained only 17 GO terms. In order to filter out highly specialized GO annotations from large set of annotations, we set two filtration criteria which are the GO level and the modified Fischer exact *p*-value to compute the enrichments and we picked only those annotations that showed meaningful GO level **(= 3)** and *p*-value **(= 0.05)**. After this filtration, we obtained 49, 23 GO terms unique to neuronal and non neuronal target matrix respectively. For common probesets, we do not apply any filtration criteria because of the small number of probesets (Figure 36).

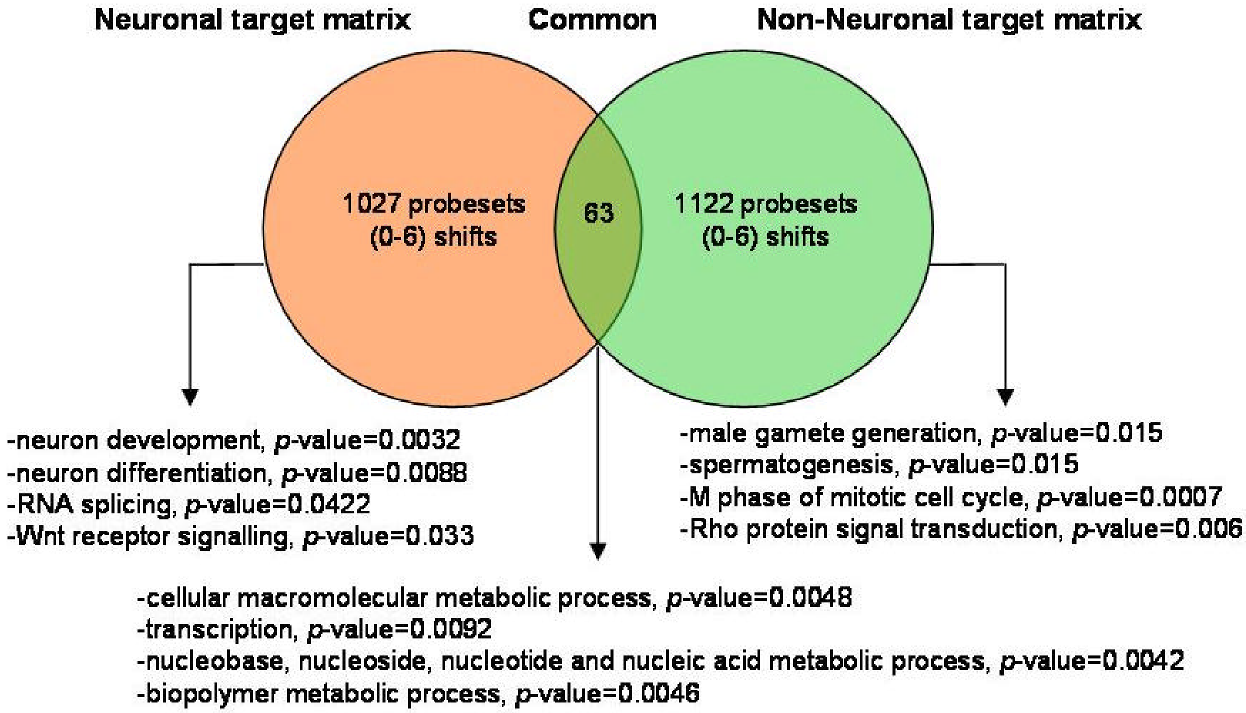

Neuronal target matrix
Common
Non-Neuronal target matrix
1027 probesets (0-6) shifts
63
1122 probesets (0-6) shifts
-neuron development, p-value=0.0032
-neuron differentiation, p-value=0.0088
-RNA splicing, p-value=0.0422
-Wnt receptor signalling, p-value=0.033
-male gamete generation, p-value=0.015
-spermatogenesis, p-value=0.015
-M phase of mitotic cell cycle, p-value=0.0007
-Rho protein signal transduction, p-value=0.006
-cellular macromolecular metabolic process, p-value=0.0048
-transcription, p-value=0.0092
-nucleobase, nucleoside, nucleotide and nucleic acid metabolic process, p-value=0.0042
-biopolymer metabolic process, p-value=0.0046

Figure 36. Multi-dimensional fitting (MDF) probesets characterization using GO analysis.
Indicating number of probesets in each of the target matrices showing (0-6) shifts with major GO annotations and their corresponding *p*-values**.**

9.4.1.1 Enriched GO terms in neuronal target matrix

GO annotations that are highly characteristic of this group (0-6 shifts) in the neuronal target matrix are neuron development (GO:0048666, GO level=7, *p*-value=0.0032, number of probesets=19), neuron differentiation (GO:0030182, GO level=6, *p*-value=0.0088, number of probesets=23), neuron morphogenesis during differentiation (GO:0048667, GO level=5, *p*-value=0.0093, number of probesets=14), Wnt receptor signaling pathway (GO:0016055, GO level=6, *p*-value=0.033, number of probesets=6).

9.4.1.2 Enriched GO terms in non-neuronal target matrix

GO annotations that are highly characteristic of this group (0-6 shifts) in the non-neuronal target matrix are male gamete generation (GO:0048232, GO level=5, *p*-value=0.015, number of probesets=17), spermatogenesis (GO:0007283, GO level=6, *p*-value=0.015, number of probesets=17), M phase of mitotic cell cycle (GO:0000279, GO level=8, *p*-value=0.00071, number of probesets=14), mitosis (GO:0007067, GO level=8, *p*-value=0.00068, number of probesets=14), and small GTPase mediated signal transduction (GO: 0007264, GO level=6, *p*-value=0.006, number of probesets=22).

9.4.1.3 Enriched GO terms in common criteria

GO annotations that are characteristic for the probesets that are common to both neuronal and non-neuronal targets are mostly general biological processes that are required for basic and normal functioning of the cells. These general biological processes are cellular macromolecular metabolic process (GO: 0044260, GO level=4, *p*-value=0.0048, number of probesets=17), transcription (GO: 0006350, GO level=5, *p*-value=0.0092, number of probesets=10), nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (GO: 0006139, GO level=4, *p*-value=0.0042, number of probesets=14) and biopolymer metabolic process (GO: 0043283, GO level=4, *p*-value=0.0046, number of probesets=8).

9.5 Cross-checking results with simple classical method

In order to cross check whether GO terms obtained through MDF analysis could also be found by simple classical method such as one-sample Student tests, we applied one-sample Students tests on datasets of both the target matrices to detect significantly up or down regulated genes and obtained respectively 35 and 4 probesets that showed less than 5% false discovery rate. Further GO analysis was performed in order to identify any common GO term between this method and those identified by MDF, but none of the GO terms were identified.

9.6 Conclusion

We have presented the Multi-Dimensional Fitting (MDF) method that simplifies different matrices of heterogeneous data describing the same group of coordinates by a mean square objective function.

The functions of MDF are shown by testing this newly developed method on transcriptomic data. We chose two target matrices, one (neuronal target matrix) related to the reference matrix (brain data) and the other (non-neuronal target matrix) one unrelated. All these datasets were treated with MDF and the results obtained were grouped based upon the number of shifts to 0. Further, we performed functional analysis using DAVID and AmiGO on the probesets that showed only 0-6 shifts. MDF performed well on these datasets, we found enrichment of neuronal related GO terms (neuron development, neuron differentiation etc) in neuronal target matrix. On the contrary, in non-neuronal target matrix where the datasets are from a variety of tissues other than brain or retina, such as liver, testis or muscle, we observed an enrichment with spermatogenesis, gamete formation and mitosis GO terms. The strong presence of spermatogenesis might be due to the presence of datasets from testis in this target matrix. The probesets that are common showed more generalized functions such as cellular macromolecular metabolic process, transcription, and biopolymer metabolic process.

We found genes characterized by RNA splicing GO annotation common to neuronal target matrix and reference matrix. According to the analysis performed on the initial reference matrix data by Mackiewcz *et al.* and dedicated to the identification of genes with temporal changes in expression during spontaneous sleep and sleep deprivation in the mouse cerebral cortex and in hypothalamus, the genes with strong differential expression exhibiting RNA splicing GO annotation are the RNA binding protein (*Rbm3*) and RNA splicing factors (*sfrs1, sfrs6, sfrs7, Fusip1, Prpf4b*). Nevertheless, after MDF analysis, additional genes that are involved in RNA splicing (*Sf3b1, Rbm4, Dbr1, Sfrs3, Sf1, Nol3, Wbp11, Snrpa, Ptbp1 and Phf5a*) could be identified. This result strongly pinpoints that the MDF analysis not only identifies the genes that are differentially regulated in the compared matrices but also genes that had few shifts compared with the overall distances computed in the two matrices and that might thus correspond to subtly regulated genes.

Finally, MDF plays a major role in noise reduction and also in elucidating the related events between two large heterogeneous data matrices which otherwise cannot be illustrated using simple methods.

9.7 Discussion

Analyzing the results obtained by MDF was a continuous learning process for us. In order to investigate the biological relevance of MDF on heterogeneous data we chose datasets for target matrices from two extremely different sources. In this analysis, a brain related sleep deprivation experiment served as the reference matrix, with data from neuronal tissues and from non-neuronal tissues (lung, muscle, testis etc) serving as the target matrices. Considering that the only coherence between the datasets of the two target matrices was the type of chip used (Affymetrix Mouse Genome array 430A 2.0) and the normalization software (RMA), MDF yielded very interesting results. Analysis through MDF detected GO annotations that were highly specific to individual target matrices, such as neuronal differentiation and development in the case of neuronal target matrix, male gamete generation and spermatogenesis for the non-neuronal matrix, and general biological process, such as cellular macromolecular metabolic process and transcription in common for both the target matrices. Such meta-analysis approaches allow searching for much broader importance and impact of differentially regulated genes in the mouse-model of sleep deprivation reference matrix.

Several RNA-splicing associated GO-categories were identified in the 0-6 shift group highlighting that this is a phenomenon of transversal importance in all the neuronal related experiments included in this study. The complexity of central nervous system assembly and function, as well as its remarkable capacity for plasticity, imply that an intricate set of mechanisms might be necessary to regulate gene expression in the brain. Alternative splicing provides a plausible means for generating diversity during development and plasticity in the nervous system (Dredge *et al.*, 2001).

Publication n^{o} 2

Multidimensional fitting for data analysis

Claude Berge [*], Nicolas Froloff [†], Ravi Kiran Reddy Kalathur [‡], Myriam Maumy [§], Olivier Poch [‡] , Wolfgang Raffelsberger [‡] , Nicolas Wicker [‡]

[*]F. Hoffmann-La Roche Ltd, Malzgasse 30 / CH-4070 / Basel, Switzerland,[‡]Laboratoire de Bioinformatique et Génomique Intégratives, Institut de Génétique et de Biologie Moléculaire et Cellulaire, 1, rue Laurent Fries, BP 10142, 67404 Illkirch Cedex, France,[§]IRMA, 7, rue René-Descartes, 67084 Strasbourg Cedex, France, and [†]CEREP, 19 avenue du Québec, 91951 Courtaboeuf Cedex, France

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Large multidimensional data matrices are frequent in domains such as biology, astronomy and economy. However, statistical methods often have difficulties to efficiently deal with such matrices because they contain very complex data. Consequently variable selection and dimension reduction methods are often used to reduce matrix complexity, although at the expense of information conservation. A new class of methods is presented for the case where two matrices are available to describe the same population. The presented method transforms with some constraints one of the matrices, called the target matrix to make it fit with the second matrix, a reference matrix which is assumed to be accurate. The fitting is done on the distances computed for the two matrices and the transformation can be as simple as a set of shifts on some of the matrix coordinates. Although the main goal is to simplify matrices, this transformation can also reveal points sharing common behaviour in the two matrices, which would have been difficult to detect otherwise. The multidimensional fitting is applied to artificial data and then to a real case problem of transcriptomic meta-analysis proving that the matrix transformation can disclose interesting relationships between complex heterogeneous datasets.

multivariate data analysis | large dimension problem | meta-analysis

Introduction

Multivariate data analysis is a rich area [4, 34, 37] with a long list of well-known methods that are used to simplify matrices or to explain existing links between numerous variables. One of the desired simplifications is dimension reduction in cases where the number of dimensions of a dataset is too large. For dimension reduction, Principal Components Analysis (PCA) has for a long time been the canonical method [24] although it is only suitable for Gaussian variables. More recently Independent Component Analysis (ICA) [27] has been introduced to deal with non-Gaussian variables. Matrix simplification is also of interest when data are described by a large number of distances. Indeed the distances are usually converted to coordinates and at the same time the dimension is reduced. Multidimensional Scaling (MDS) [34] is similar to PCA in that it maximizes the inertia obtained on each axis. Another possible simplification is variable selection [26] which eliminates non informative variables.

Another important application of multivariate analysis is to explain data with other data. In particular, when two data matrices are available, one classical objective is to explain one matrix with the other through a Generalized Linear Model (GLM) [33], which provides a large class of tools to extend the linear regression. Another way of studying the dependence of the matrices is to perform a Canonical Correlation Analysis [34] (CCA) which at the same time reduces the dimensionality. Indeed, CCA establishes the best correlation between linear combinations of the two sets of variables. Finally, all these methods can be extended using the kernel trick [2] which implicitly projects the data sets in an unknown Hilbert space.

The key idea we introduce in this paper is that a large dimensional data matrix can be transformed to obtain more accurate data if another matrix is provided that contains accurate data. These two matrices will be called respectively the target and reference matrix. Under specified constraints, the presented method transforms values of the target matrix in such a way that the distances computed on the target matrix are as close as possible to the distances present in the reference matrix. The constraints are chosen according to the type of problem that is dealt with. Potentially all possible transformations and error measures can be devised. Henceforth these methods will be referred to as Multidimensional Fitting (MDF) owing to the type of transformation that is performed. The MDF method is first tested on artificial data, then its usefulness is illustrated by performing meta-analysis of gene expression data from independent microarray experiments to mine and identify biological features present both in the target and reference matrices.

Multidimensional Fitting

Let us consider n points described by two sets of variables. The first set is composed of p variables and the second of q, not necessarily equal to p. Our objective is to modify the first matrix so that the distances computed on this matrix are as close as possible to the distances d_{ij} computed on the second matrix for each pair of points i and j. In this presentation, the function to optimize is the mean square error:

$$E(a,\theta) = \sum_{1\le i<j\le n} \left(d_{ij} - ad(f(x_i,\theta), f(x_j,\theta))\right)^2 \qquad \textbf{[1]}$$

where f is an arbitrary function that is specified depending on the problem at hand, θ is the function parameter to optimize and a a scaling variable to adjust the target matrix with the reference matrix.

For the artificial example that will be considered, formula 1 becomes:

$$E(a,\mathbf{y}) = \sum_{1\le i<j\le n} \left[d_{ij} - a\sqrt{\sum_{k=1}^{p} \left[(x_{ik}+y_{ik}) - (x_{jk}+y_{jk})\right]^2} \right]^2$$

with the constraints that $-x_{ik} \le y_{ik} \le 0$ and $a \ge 0$

where x_{ik} is the k^{th} coordinate of point i and y_{ik} the corresponding variable to optimize. Intuitively, the y_{ik} can decrease the value of a coordinate to bring point i close to other points. One can note that if all the y_{ik} are known, a can be expressed in the following closed

This paper was submitted directly to the PNAS office.

form:

$$\frac{\partial E}{\partial a} = 2 \sum_{1 \leq i < j \leq n} \left(- \sqrt{\sum_{k=1}^{p} \left[(x_{ik} + y_{ik}) - (x_{jk} + y_{jk}) \right]^2} \right)$$

$$\left(d_{ij} - a \sqrt{\sum_{k=1}^{p} \left[(x_{ik} + y_{ik}) - (x_{jk} + y_{jk}) \right]^2} \right) = 0$$

$$\Rightarrow a = \frac{\sum_{1 \leq i < j \leq n} d_{ij} \sqrt{\sum_{k=1}^{p} \left[(x_{ik} + y_{ik}) - (x_{jk} + y_{jk}) \right]^2}}{\sum_{1 \leq i < j \leq n} \sum_{k=1}^{p} \left[(x_{ik} + y_{ik}) - (x_{jk} + y_{jk}) \right]^2} \quad [2]$$

Discrete multidimensional fitting. The problem can be simplified by considering the discrete case, where each y_{ij} can only be equal either to 0 or to $-x_{ij}$. A simple example is presented in Fig. 1 where 3 different target triangles are transformed with either 0, 1 or 2 coordinates shifted to 0 and optionally a scaling to a reference triangle. This problem is NP-complete since it is a more general problem than binary multidimensional scaling which has been proven to be NP-complete [19]. To optimize the above function, a relaxation method has been used. Each coordinate has been optimized iteratively. After each iteration, a is computed according to formula 2.

Continuous multidimensional fitting. In the continuous case, the Gibbs sampler [23] can be applied that incorporates in a classical way the function to optimize into a probability distribution $p(y|x)$ by setting $p(y|x) = A \exp\{-E(y)\}$ where A is the normalization constant [28]. In each iteration t of the Gibbs sampler there are $n \times p$ steps since there $n \times p$ univariate variables to sample. In our case, variable y_{ij}^t is sampled from the conditional distribution given all the other variables at their current values:

$$p(y_{ij}^t | y_{kl}^t, y_{mn}^{t-1}) \text{ with } kl \prec ij \text{ and } ij \prec mn$$

where $\prec$ is defined in the following way: $kl \prec ij$ iif $k < i$ or, $k = i$ and $l < j$. This conditional distribution is sampled using the Metropolis algorithm [35] using, for each y_{ij}, the uniform distribution on the interval $[-x_{ij}, 0]$.

Simulated dataset

Both the discrete and the continuous methods have been tested on 10 datasets. In each case a matrix noted M_1 has been simulated randomly that consists of a number of points ranging from 100 to 1000 with 20 coordinates. Each coordinate has been generated randomly between 0 and 100 according to the uniform law. Next a distance matrix D is computed on M_1. Then, a new matrix, M_2, is defined by choosing randomly for each point, 10 coordinates that are set to 0 so that M_2 contains 1000 to 10000 errors compared to M_1. The optimizations are performed on M_2 and D with 20, 100 or 500 iterations producing a new matrix denoted R. R is finally compared to M_1. In each test, the discrete method discovered the optimum matrix after 20 iterations whereas the Gibbs sampling did not. Results of the latter are all reproduced on table 1 showing that they greatly improve between the 20^{th} and the 500^{th} iteration. These simulations show that when a discrete result is expected it is much faster to use the discrete MDF than the Gibbs sampling which should be reserved for relatively small matrices in the continuous case.

Application to a biological dataset

We have used the multidimensional fitting method to illustrate its usefulness in the case of meta-analysis of heterogeneous transcriptomic data. The goal of this study was to identify genes showing similar expression behaviour in two sets of experiments. The expression data of a gene is represented by one or more probe sets for which measures are obtained from different experiments. These experiments can contain one or more data samples, corresponding to different experimental conditions. The optimization method that was used was the discrete MDF as it gave very good results on the artificial data and as the matrices size are too large for the continuous MDF.

Datasets used for the analysis. Two target matrices of 22960 points (probe sets) in 35 dimensions (samples) were constructed (table 2). The 2 distinct target matrices are composed of 35 microarray samples each, from neuronal (brain and retina) and non-neuronal tissues (mainly muscle and lung). The reference matrix has also 22960 points and is composed of 16 datasets (16 dimensions), from an experiment studying the effects of spontaneous sleep and prolonged wakefulness on the mouse brain (table 2). Thus the data used were very heterogeneous, coming from different experiments done in different laboratories with different protocols.

All datasets were pre-processed using standard treatments. First the Robust Multi-array Average (RMA) algorithm [9] was applied, next log2 gene expression ratios were determined (using corresponding control samples as denominator) and then normalized using the quantile-quantile method. All procedures were performed in R using the packages from Bioconductor [22]. The MDF was applied twice, once on the neuronal target matrix and once on the non-neuronal target matrix using the same brain reference matrix resulting in two different transformed target matrices.

Results. For each transformed target matrix and each sample, the number of probe sets having a shift to 0 is represented in figure 2. It is distributed homogeneously ranging from 13018 to 17687 (out of 22960) probe sets. The number of probe sets having a shift represents approximately 64% of each sample in the neuronal as well as in the non-neuronal target matrix.

The cumulative curves of the number of probe sets compared to the number of shifts are similar for both targets. (Fig. 3). However the cumulative curve of the common probe sets is significantly shifted to the right compared to the other two curves. This means that there are only a few probe sets with a small number of shifts that are common to the two target matrices. To identify these probe sets we considered the histograms corresponding to the target matrix curves. The frequency distribution indicates, for both target matrices, a bimodal distribution with a main mode centered around 26 shifts (Fig. 4) covering approximately 85% of all probe sets and a small mode for the probe sets having 0 to 6 shifts.

In order to get an overview of the biological and molecular processes of the genes corresponding to the groups of probe sets characterized by 0 to 6 shifts for the two target matrices, we used a standard ontology used in biology called the Gene Ontology (GO) [25]. GO terms are attributed to the genes using the DAVID software [18] and their enrichment, i.e significant presence of a given GO term in a group, is computed using the modified Fisher exact test.

The two groups obtained for the neuronal and the non-neuronal target matrices share only 63, out of 1090 and 1185 probe sets respectively, which are mainly involved in general cell functions such as cell organisation and primary metabolism. For 0 to 6 shifts, we observed GO annotations that were highly characteristic to each group. In particular, in the case of the neuronal target matrix this group was very enriched in highly specialised molecular functions previously associated with brain (gene-ontology: RNA splicing (GO:0008380), RNA splicing, via transesterification reactions (GO:0000398), neuron morphogenesis during differentiation (GO:0048667), neuron development (GO:008380), neuron differentiation (GO:0030182), nervous system development (GO:0007399), chromatin remodelling (GO:0006338)).

In contrast, the group of probe sets having 0 to 6 shifts from the non-neuronal target matrix are primarily composed of general GO terms such as regulation of cellular physiological process (GO:0050794), regulation of metabolism (GO:0019222) or macromolecule metabolism (GO:0043170). In this group the only GO terms that were found to be enriched were M phase of mitotic cell cycle (GO:0000087) and small GTPase mediated signal transduction (GO:0007264).

As a control we checked that the probe sets having a small number of shifts were not simply those that had a strong under or overexpression. Therefore, one-sample Student tests were performed on all the datasets to detect probe sets with fold changes significantly different to 0. Then, probe sets that were under 5% false discovery rate [8] were selected and only 2 of them were present in the group of probe sets having 0 to 6 shifts in the neuronal target matrix, whereas none were present for the non-neuronal target matrix.

Conclusion and perspectives

We have presented a new class of methods called Multidimensional Fitting (MDF) for simplifying matrices of data describing the same population of points. This method fits a target matrix to a reference matrix by transforming it under some predefined constraints, resulting in an optimization problem where the objective function is typically a mean square error. An example of the method has been shown in the case of a simple translation transformation on artificial datasets with a mean square error. Then, as a biological application, two target matrices, one with neuronal tissue and the other with non-neuronal tissue gene expressions were treated by MDF together with a brain reference matrix. This example was a challenging one due to the heterogeneity of the datasets composing each matrix which could have biased the results. The MDF performed well in this test in that it found neuronal related genes for the neuronal target matrix and very general genes for the non-neuronal matrix when fitted to a brain reference matrix.

The MDF could be further developed by introducing more complex transformations than coordinate translations and by changing the objective function. For the case we considered, euclidean distances between points were fitted, but these distances could be replaced by any relevant distance or proximity measure. The optimization methods that were tested included the relaxation and the Gibbs algorithm for the discrete and the continuous cases respectively. However, owing to the complexity of the objective function it would be interesting to test other optimization methods, which could perform better, for example simulated annealing or genetic algorithms.

The transformation could also be modified to reflect some constraints on the data, if for example the points have standardized coordinates (mean$=0$ and standard deviation$=1$) it is necessary to keep the new coordinates standardized. Furthermore, as mentioned above it is always possible to consider either a discrete or a continuous transformation. The discrete MDF is suitable for instance for data where it is known that some values contain no signal and only noise. The desired transformation for these values would be to set them to a constant value corresponding to a total absence of signal, typically 0. This value is essential to compute distances on the target matrix that are distributed in the same way as the distances of the reference matrix. The discrete case is also interesting, as shown in this paper, to speed up the optimization which can be rather slow in the continuous case.

Finally we believe that MDF will prove valuable in a wide range of applications where complex datasets are available and difficult to use directly because of noise or hidden relationships. Such applications can be for example prediction of adverse events for a set of drugs with in vitro assay data available [29] or preference mapping where the goal is to explain consumer preferences for a set of competitive products by their sensory profiles for these products [15]. With such and a variety of other cases MDF can be combined with subsequent supervised or unsupervised learning, for which MDF can highlight the values that are crucial for the learning process.

This work was supported by Cerep, the European Retinal Research Training Network (RETNET) MRTN-CT-2003-504003, CNRS, INSERM and the Université Louis Pasteur. We are also grateful to Julie Thompson for critical reading of the manuscript.

1. Abou-Sleymane, G., Chalmel, F., Helmlinger, D., Lardenois, A., Thibault, C., Weber, C., Merienne, K., Mandel, J.L., Poch, O., Devys, D. and Trottier, Y. (2006) *Hum Mol Genet, 15*, 691-703.
2. Aizerman, M., Braverman, E., and Rozonoer, L. (1964) *Automation and Remote Control, 25*, 821-837.
3. Akimoto, M., Cheng, H., Zhu, D., Brzezinski, J. A., Khanna, R., Filippova, E., Oh, E. C., Jing, Y., Linares, J. L., Brooks, M., Zareparsi, S., Mears, A. J., Hero, A., Glaser, T. and Swaroop, A. (2006) *Proc Natl Acad Sci U S A, 103*, 3890-3895.
4. Anderson, T.W. (2003) An Introduction to Multivariate Statistical Analysis, *Wiley-Interscience*.
5. Bach, F.R., and Jordan, M.I. (2002) *Machine Learning Research, 3*, 1-48.
6. Baker, P. E., Kearney, J. A., Gong, B., Merriam, A. P., Kuhn, D. E., Porter, J. D. and Rafael-Fortney, J. A. (2006) *Neurogenetics, 7*, 81-91.
7. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., and Edgar, R. (2007) *Nucleic Acids Research, 35*.
8. Benjamini, Y. and Hochberg, Y. (1995) *J.R.S.S. Series B(Methodological), 57*, 289-300.
9. Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P (2003) *Bioinformatics, 19*, 185-93.
10. Bray, J.H. and Maxwell S.E. (1985) in Multivariate Analysis of Variance, *Sage Publications*.
11. Cao, Y., Kumar, R. M., Penn, B. H., Berkes, C. A., Kooperberg, C., Boyer, L. A., Young, R. A. and Tapscott, S. J. (2006) *Embo J, 25*, 502-11.
12. Cheng, H., Aleman, T. S., Cideciyan, A. V., Khanna, R., Jacobson, S. G. and Swaroop, A. (2006) *Hum Mol Genet, 15*, 2588-602.
13. Christensen, R. (1997) in Log-Linear Models and Logistic Regression, *Springer*.
14. Cottet, S., Michaut, L., Boisset, G., Schlecht, U., Gehring, W. and Schorderet, D. F. (2006) *Faseb J, 20*, 2036-49.
15. Courcoux, P.h., and Chavanne, P.C. (2001) *Preference mapFood Quality and Preference, 12*, 369-372.
16. D'Souza, I. and Schellenberg, G.D. (2000) *J Biol Chem 275*, 17700-9.
17. Denolet, E., De Gendt, K., Allemeersch, J., Engelen, K., Marchal, K., Van Hummelen, P., Tan, K. A., Sharpe, R. M., Saunders, P. T., Swinnen, J. V. and Verhoeven, G. (2006) *Mol Endocrinol, 20*, 321-334.
18. Dennis, G.Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003) *Genome Biology, 4(5)*.
19. Deza, M.M., and Laurent, M. (1997)in Geometry of Cuts and Metrics, *Springer*.
20. Edwards, M.G., R. M. A., Yuan, M., Kendziorski, C.M., Weindruch, R., and Prolla, T.A. (2007) *BMC Genomics, 8(80)*.
21. Garey, M.R., and Johnson,D.S. (1979) Computers and Intractability: A Guide to the Theory of NP-Completeness, *W.H. Freeman*.
22. Gentleman, R.C, Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., and Zhang, J. (2004) *Genome Biol., 5(10)*.
23. Geman, S., and Geman, D. (1984) *IEEE T. Pattern Anal., 6* 721-741.
24. Greenacre, M.J. (1984) in Theory and Applications of Correspondence Analysis, *Academic Press*.
25. Harris, H.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., and White, R. (2004) *Nucleic Acids Research, 32* D258-D261.
26. Hocking, R.R. (1976) *Biometrics, 32* 1-49.
27. Jutten, C., and Hérault, J. (1991) *Signal Processing, 24* 1-10.
28. Schneider, J.J., and Kirkpatrick, S. (2006) in Stochastic Optimization, *Springer*.
29. Krejsa, C.M., Horvath, D., Rogalski, S.L., Penzotti, J.E., Mao, B., Barbosa, F., and Migeon, J.C. (2003) *Curr Opin Drug Discov Devel., 6(4)* 470-480.

30. Li, J., Grigoryev, D. N., Ye, S. Q., Thorne, L., Schwartz, A. R., Smith, P. L., O'Donnell, C. P. and Polotsky, V. Y. (2005) *J Appl Physiol*, *99*, 1643-1648.

31. Lin, C.L., Bristol, L. A., Jin, L., Dykes-Hoberg, M., Crawford, T., Clawson, L. and Rothstein, J.D. (1998) *Neuron 20* 589-602.

32. Majumder, P. K., Febbo, P. G., Bikoff, R., Berger, R., Xue, Q., McMahon, L. M., Manola, J., Brugarolas, J., McDonnell, T. J., Golub, T. R., Loda, M., Lane, H. A. and Sellers, W. R. (2004) *Nat Med*, *10*, 594-601.

33. McCullagh, P. and Nelder J. (1989) in Generalized Linear Models, *Chapman and Hall*.

34. Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) in Multivariate Analysis, *Academic Press*.

35. Metropolis, N., Rosenbluth, M.N., Rosenbluth, A.H., Teller, A.H. and Teller, E. (1953) *J. Chem. Phys.*, *21(6)*, 1087-1092.

36. Meyer, T., Fromm, A., Munch, C., Schwalenstocker, B., Fray, A.E., Ince, P.G., Stamm, S., Gron, G., Ludolph, A.C. and Shaw, P.J. (1999) *J Neurol Sci 170* 45-50.

37. Muirhead, R.J. (2005) Aspects of Multivariate Statistical Theory, *Wiley-Interscience*.

38. Nilsson, E. C., Long, Y. C., Martinsson, S., Glund, S., Garcia-Roves, P., Svensson, L. T., Andersson, L., Zierath, J. R. and Mahlapuu, M. (2006) *J Biol Chem*, *281*, 7244-7252.

39. Okubo, T., Knoepfler, P. S., Eisenman, R. N. and Hogan, B. L. (2005) *Development*, *132*, 1363-74.

40. Parker, G. E., Pederson, B. A., Obayashi, M., Schroeder, J. M., Harris, R. A. and Roach, P. J. (2006) *Biochem J*, *395*, 137-145.

41. Semeralul, M.O., Boutros, P.C., Likhodi, O., Okey, A.B., Van Tol, H.H. and Wong, A.H. (2006) *J Neurobiol 66*, 1646-58.

42. Steele, M.R., Inman, D.M., Calkins, D.J., Horner, P.J. and Vetter, M.L. (2006) *Invest Ophthalmol Vis Sci 47*, 977-85.

Figures

Figure 1 The largest triangle in full line is the reference triangle. The dotted triangle (0) needs no coordinate shift to 0 but only a scaling. The dashed-line triangle (1) needs one shift to 0 of a y coordinate and a scaling. The dotted dashed-line triangle (2) needs 2 shifts to 0 and no scaling.

Figure 2 Number of probe sets that have at least one shift to zero after MDF. X-axis indicates the sample number and Y-axis indicates the number of shifted probe sets.

Figure 3 The cumulative curves of the number of probe sets respective to the number of shifts are plotted for the neuronal and the non-neuronal target matrices. The cumulative curve for the common probe sets, i.e. those having a number of shifts inferior or equal to a given abscissa, is also shown.

Figure 4 After MDF, the number of probe sets per number of shifts is plotted as a histogram for the neuronal and the non-neuronal target matrices. The graph indicates that approximately 85% of the probe sets form a bell shaped curve. The group of probe sets that underwent 0-6 shifts in one of the target matrices was selected for further analysis.

Tables

Table 1. 10 datasets have been generated to assess the performance of the Gibbs sampler to solve the continuous MDF problem. Each dataset is composed of a matrix M1 containing a number of points ranging from 100 to 1000 with 20 coordinates and a matrix M2 where 10 coordinates for each point of M1 are chosen randomly and set to 0. After 20, 100 and 500 iterations, the number of times the final coordinate is at a distance superior to 5 from the initial coordinate in M1 is reported.

number of points	100	200	300	400	500	600	700	800	900	1000
20 iterations	804	1383	2001	2583	3144	3789	4302	4853	5355	6192
100 iterations	463	285	310	322	245	264	263	334	334	384
500 iterations	150	41	24	9	4	2	2	5	4	3

Table 2. Datasets used for MDF analysis. The GEO accession number [7], the respective publication of the original data, the tissue origin and the number of original samples used for this comparison are indicated. The datasets were generated using Affymetrix GeneChip Mouse Genome 430A 2.0 (http://affymetrix.com).

GEO accession	Publication	Tissue type	Number of samples
Neuronal Target matrix			
GSE3634	Abou-Sleymane G. et al. [1]	Retina	2
GSE1816	Atul Butte et al.	Retina	4
GSE3249	Cottet S. et al. [14]	Retina	3
GSE4051	Akimoto M. et al. [3]	Retina	5
GSE1835	Ruiz A.C. et al.	Retina	1
GSE5338	Cheng H. et al. [12]	Retina	1
GSE3554	Steele M.R. et al. [42]	Retina	1
GSE1999	Salomon E. et al.	Brain	2
GSE2867	Salomon E. et al.	Brain	4
GSE2869	Salomon E. et al.	Brain	1
GSE4675	Semeralul M.O. et al. [41]	Brain	4
GSE4758	Salomon E. et al.	Brain	3
GSE4040	Salomon E. et al.	Brain	1
GSE6678	Salomon E. et al.	Brain	3
Non-Neuronal Target matrix			
GSE4067	Nilsson E.C. et al. [38]	Muscle	2
GSE5304	Hoffman E. et al.	Muscle	4
GSE1303	Hoffman E. et al.	Lung	4
GSE6323	Edwards M.G. et al. [20]	Muscle	2
GSE2198	Pederson B.A. et al.	Muscle	4
GSE6077	Cox B. et al.	Lung	1
GSE1471	Kaminski H.J. et al.	Cardiac muscle	2
GSE1873	Li J. et al. [30]	Liver	1
GSE3858	Cao Y. et al. [11]	Skeletal muscle	7
GSE1413	Majumder P.K. et al. [32]	Prostate	2
GSE4711	Iguchi N. et al.	Testis	2
GSE2259	Denolet et al. [17]	Testis	4
Reference matrix			
GSE6514	Mackiewicz M. et al.	Brain	16

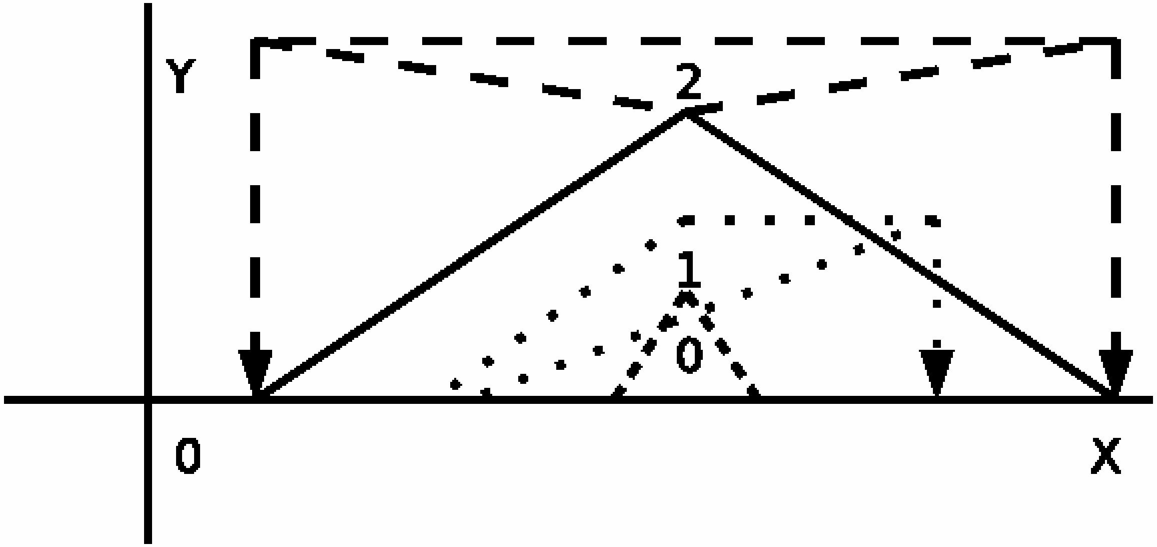


Figure 1

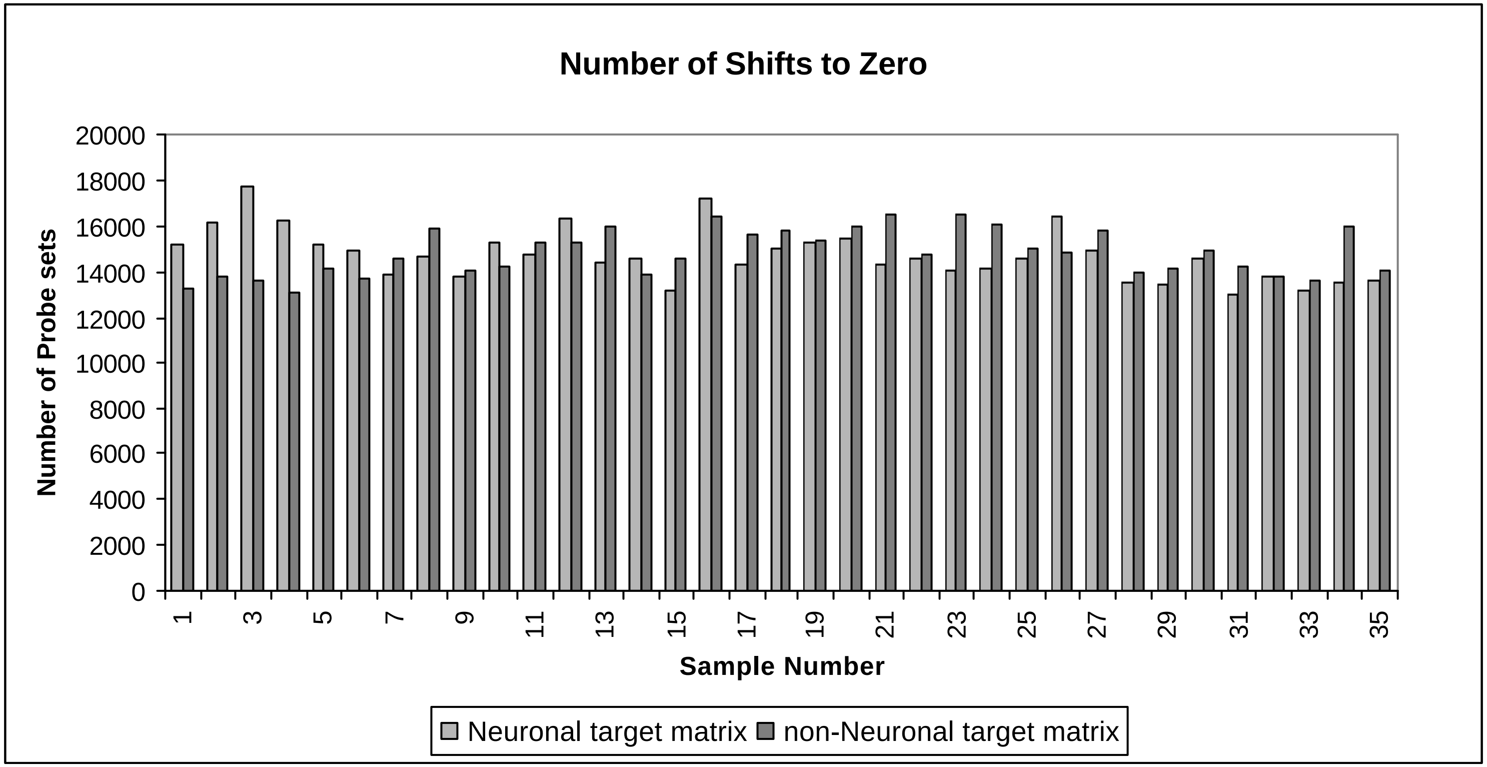

Number of Shifts to Zero
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
Number of Probe sets
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
Sample Number
Neuronal target matrix
non-Neuronal target matrix

Figure 2

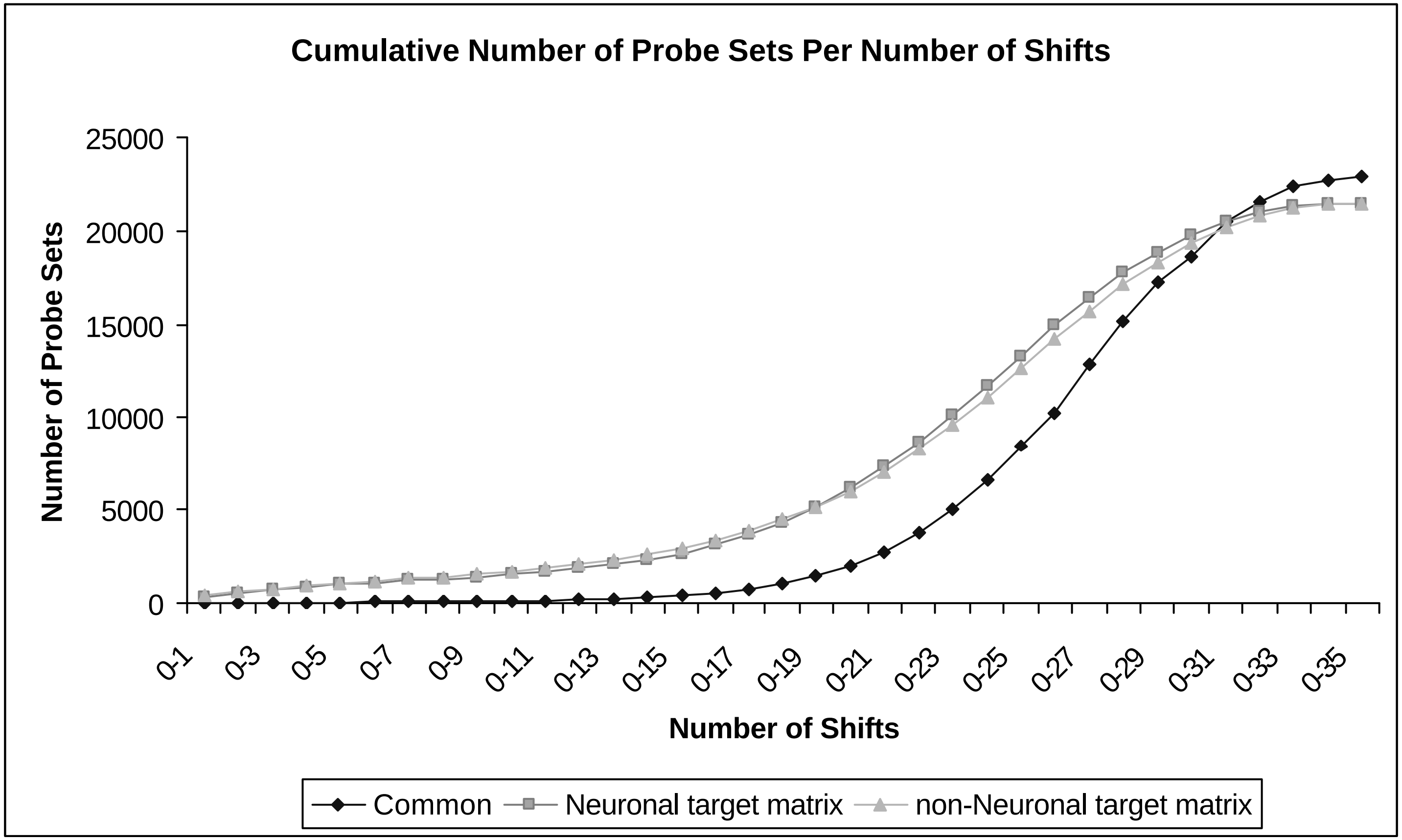

Cumulative Number of Probe Sets Per Number of Shifts
Number of Probe Sets
25000
20000
15000
10000
5000
0
0-1
0-3
0-5
0-7
0-9
0-11
0-13
0-15
0-17
0-19
0-21
0-23
0-25
0-27
0-29
0-31
0-33
0-35
Number of Shifts
Common
Neuronal target matrix
non-Neuronal target matrix

Figure 3

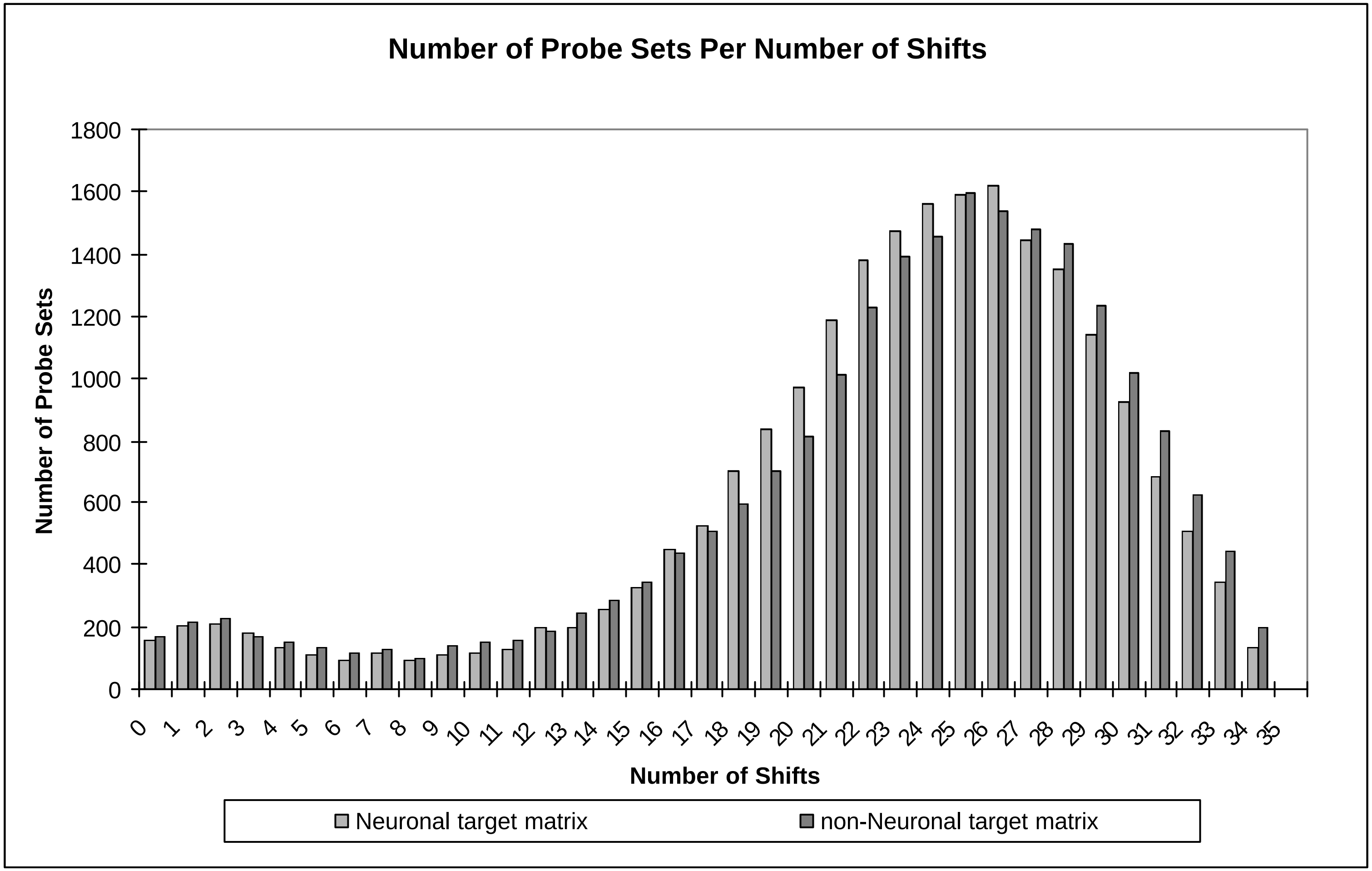
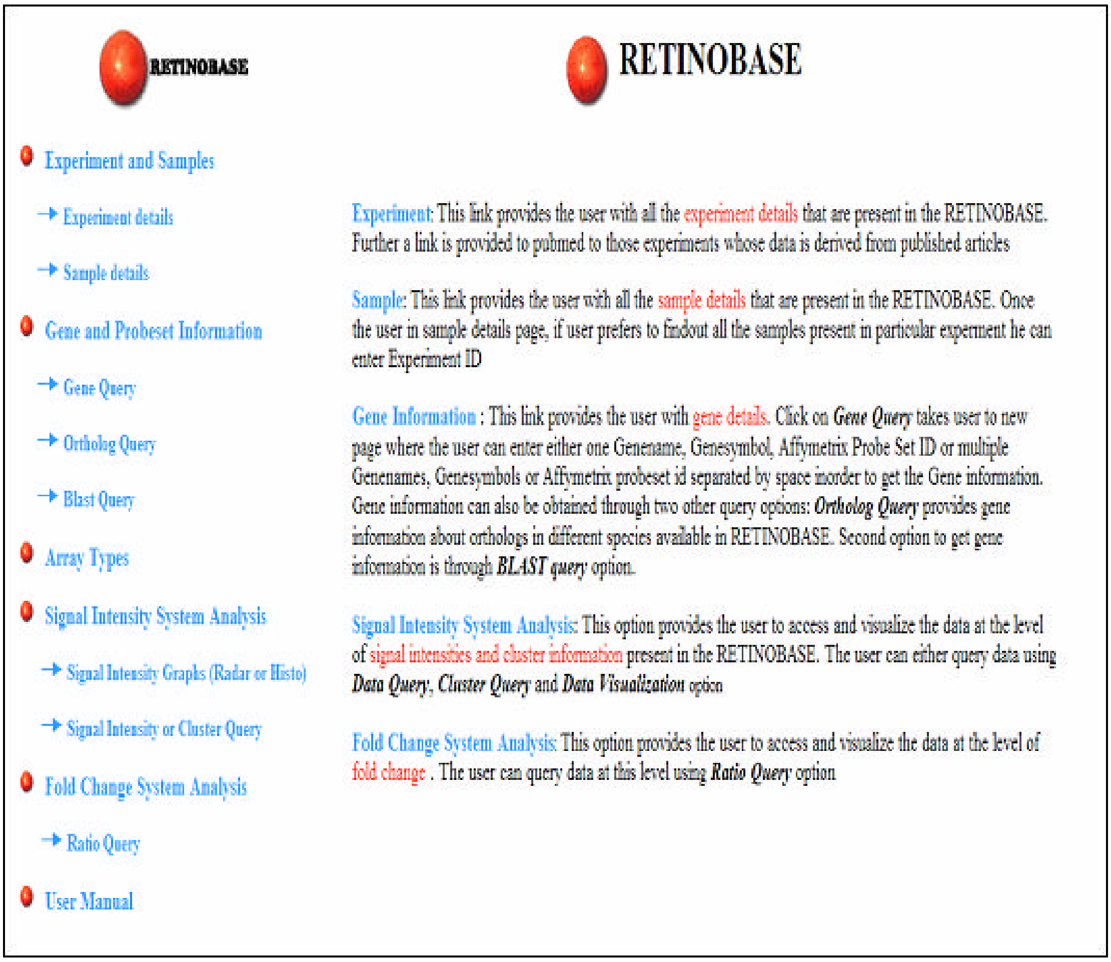

Number of Probe Sets Per Number of Shifts
Number of Probe Sets
1800
1600
1400
1200
1000
800
600
400
200
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
Number of Shifts
▨ Neuronal target matrix
■ non-Neuronal target matrix

Figure 4

Chapter 10. Architecture, data query system and data visualization aspects of gene expression database: RETINOBASE (Publication 3)

10.1 Scientific context

Gene expression studies that include high-throughput techniques such as microarrays generate large volumes of different types of data that need to be stored, treated and queried to extract useful information. In this context, we developed RETINOBASE, a relational database (***http://alnitak.u-strasbg.fr/RetinoBase/*.**) and its associated website (Figure 37). This project allowed us to approach all the aspects of the creation of a post-genomic relational database involving, the design, development and deployment of a database and also the mining, analysis of experiments from public gene expression data repositories and incorporation into RETINOBASE. In this chapter, we describe the basic architecture of the data storing, processing, as well as various querying systems and powerful data visualisation system available in RETINOBASE.


RETINOBASE
RETINOBASE
Experiment and Samples
Experiment details
Sample details
Gene and Probeset Information
Gene Query
Ortholog Query
Blast Query
Array Types
Signal Intensity System Analysis
Signal Intensity Graphs (Radar or Histo)
Signal Intensity or Cluster Query
Fold Change System Analysis
Ratio Query
User Manual

Experiment: This link provides the user with all the experiment details that are present in the RETINOBASE. Further a link is provided to pubmed to those experiments whose data is derived from published articles

Sample: This link provides the user with all the sample details that are present in the RETINOBASE. Once the user in sample details page, if user prefers to findout all the samples present in particular experment he can enter Experiment ID

Gene Information : This link provides the user with gene details. Click on *Gene Query* takes user to new page where the user can enter either one Genename, Genesymbol, Affymetrix Probe Set ID or multiple Genenames, Genesymbols or Affymetrix probeset id separated by space inorder to get the Gene information. Gene information can also be obtained through two other query options: *Ortholog Query* provides gene information about orthologs in different species available in RETINOBASE. Second option to get gene information is through *BLAST query* option.

Signal Intensity System Analysis: This option provides the user to access and visualize the data at the level of signal intensities and cluster information present in the RETINOBASE. The user can either query data using *Data Query*, *Cluster Query* and *Data Visualization* option

Fold Change System Analysis: This option provides the user to access and visualize the data at the level of fold change . The user can query data at this level using *Ratio Query* option

Figure 37. RETINOBASE home page (http://alnitak.u-strasbg.fr/RetinoBase/)
Indicating general information on experiment and sample details. Specific query options are accessible through the left panel.

RETINOBASE was developed in the context of the "European Retinal Research Training Network" (RETNET) an FP6 (6th Framework programme) Research Training Network (RTN). We have worked in close association with Dr. Thierry Leveillard's laboratory in Paris. The valuable discussions, inputs and suggestion from his group and members of RETNET have helped in the design, development and the setting up of our database.

10.2 Overview of a microarray experiment

A microarray experiment is performed to study the difference of gene expression at mRNA level in control and test samples. For example, to study the effects of drug on retina in mice we have two groups, a control group and test group, to which drug treatment is given. Mice in both groups are sacrificed and retina is obtained, mRNA is isolated from retinal tissues and is reverse transcribed to labelled cDNA and hybridised on to two separate Affymetrix GeneChip array. These hybridized chips are scanned and data are obtained as CEL files (contains intensity information for a given probe on an array). Then, they are analysed using various algorithms to obtain normalized data at the level of signal intensity or at the level of fold change. This normalized information associated to probeset is clustered. This data at the level of signal intensity, fold change and cluster is stored in RETINOBASE. The normalization and clustering are actually done outside RETINOBASE. We plan to integrate these two modules in RETINOBASE in future.

10.3 RETINOBASE architecture

In this section, the storage aspects of the database are first quickly described and later, the querying and visualization tools offered by the RETINOBASE web site.

We built the RETINOBASE schema (Figure 43) based on 7 different modules which are basic components of any microarray experiment. These seven different modules are organized into 26 different tables linked together within a relational database management system (RDBMS). The 7 modules are:

- **Probeset**
- **Array**
- **Organism**
- **Sample**
- **Experiment**
- **Data**
- **Cluster**

The "**Probeset**" module stores the probeset id (a unique name) and various information concerning the gene associated with the probeset such as : gene name, gene symbol, chromosomal location, OMIM, linked retinal disease information, gene ontology, etc. All this information is furnished from Affymetrix (http://www.affymetrix.com) and the retinal disease information, from RetNet (Retinal information network) (http://www.sph.uth.tmc.edu/Retnet/). Some SQL scripts were developed to automatically introduced the latter information in the database and link it with gene information.

The “**Array”** module contains details about array type, manufacturer, array name and description. At present, we have four different array types concerning mouse, and one each concerning drosophila, zebrafish, chicken, rat and human (details about array types are described in annexe 3).

As one probeset can be present in different arrays, we have to introduce a “Many to Many” relationship between “Probeset” and “Array” tables. In order to facilitate this relationship, we introduce a link table (Ln_Arraytype_probeset) illustrated in (Figure 38).

An important point is to be able to cross reference probesets between different Affymetrix GeneChips that represent different species, this is possible through the “**Ortholog**” table in combination with “Probeset” table. This helps the user to identify all the probesets that belong to the same gene from all arrays representing different species.

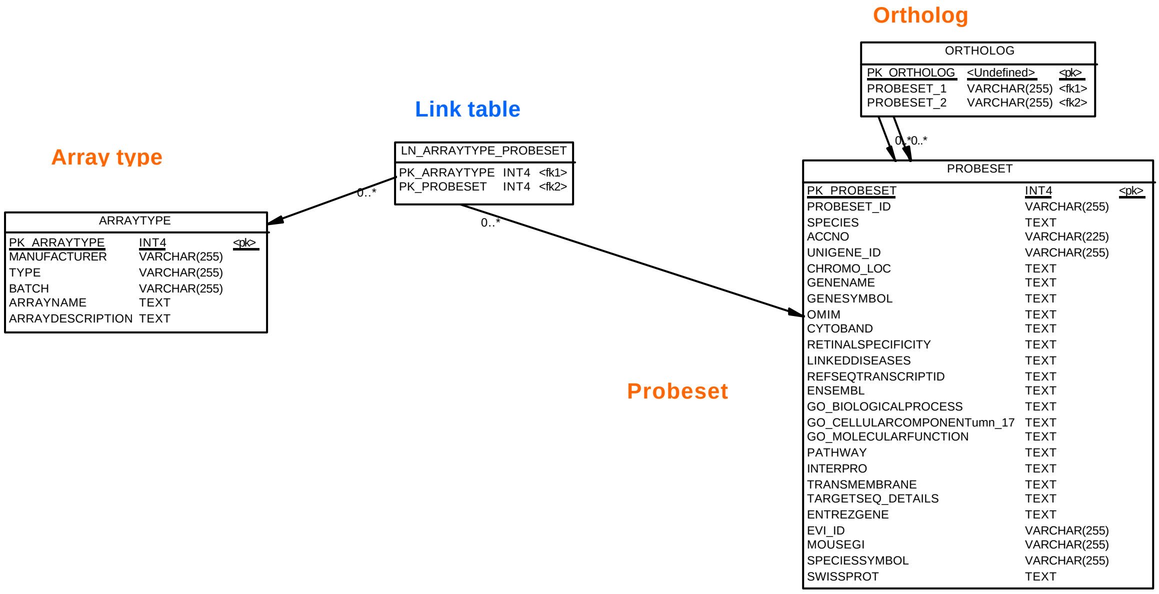


Figure 38. Schematic representation of Probeset, Array module and Ortholog modules in RETINOBASE.
Link table between probeset and array tables elucidates “Many-Many” relationship between these tables. “pk” indicates primary key, which should be unique to the table and “fk” indicate foreign key. Int, text, varchar indicates data types.

The **“Organism”** module has three different tables “Organism”, “Genotype” and “Individual” which links the first two tables. This enabled us to incorporate information concerning age, genetic background and genotype of the organism. Currently, in RETINOBASE, we have 6 different organisms (drosophila, zebrafish, chicken, rat, mouse and human), with 7 different strains in mouse and two different strains in rat.

“Sample” module has five tables “Sample”, “Sample_condition”, “Tissue”, “Treatment” and “Treatment type” linking information related to sample description, tissue type, treatment type, treatment description, which basically describes a sample in microarray experiment. Sample is derived from an organism or cell in a biological experiment, so these two modules are linked through table “Individual”. (Figure 39)

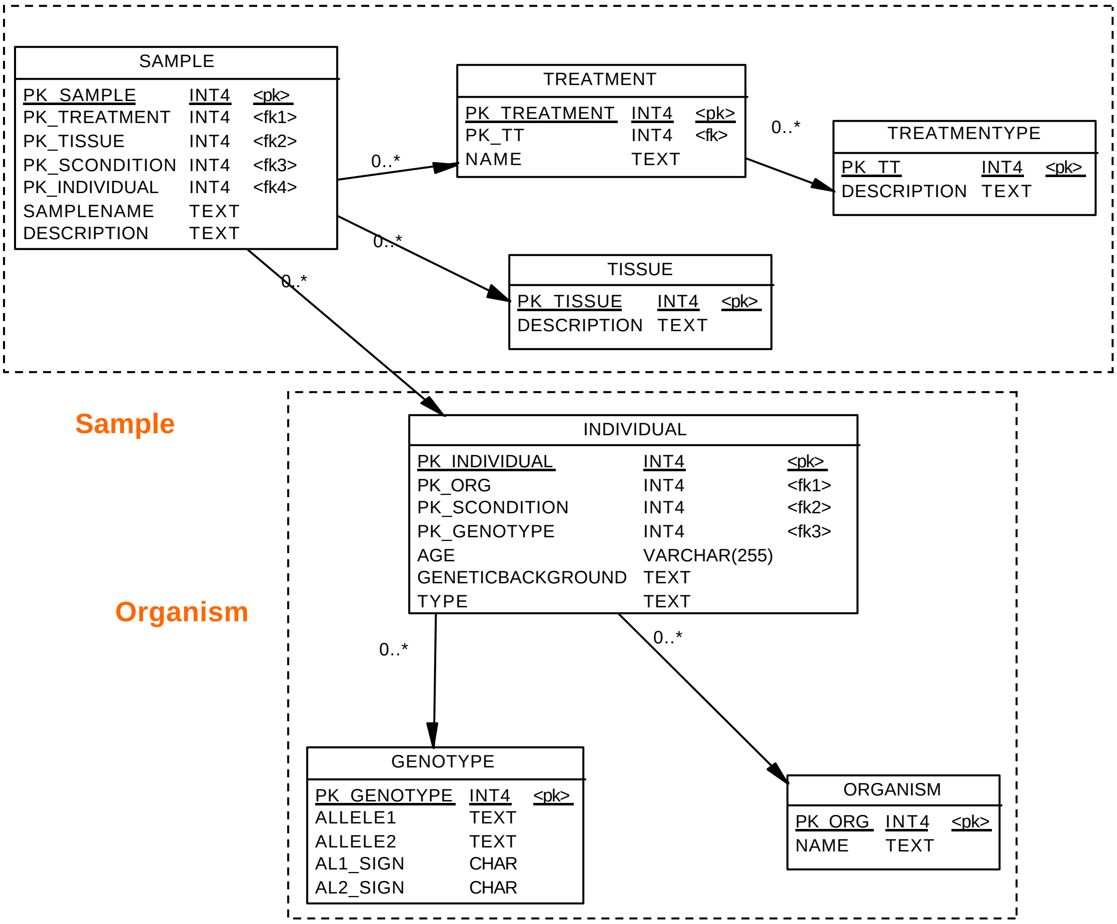


Figure 39. Schematic representation of Organism and Sample modules in RETINOBASE.

“pk” indicates primary key, which should be unique to the table and “fk” indicate foreign key. Int, text, varchar indicates data types.

The 2 modules (Array type and sample) described above and linked to "Experiment" module through a link table "Real exp" which can be correlated to one hybridization step in a microarray experiment.

Experiment module (Figure 40) provides details about experiments such as the description, the publication title, authors, abstract and quality control information (at present this is calculated outside RETINOBASE and then incorporated into the database), pubmed identifier or GEO identifier of the experiment in case of public experiments. Through pubmed identifier, users can directly access the published article in pubmed and by providing GEO identifier; we directly link the experiment to raw data source that can be downloaded from the GEO database. The information about quality control files and also various protocols such as RNA extraction, Hybridization etc corresponding to the experiment are stored in table "Expfiles".

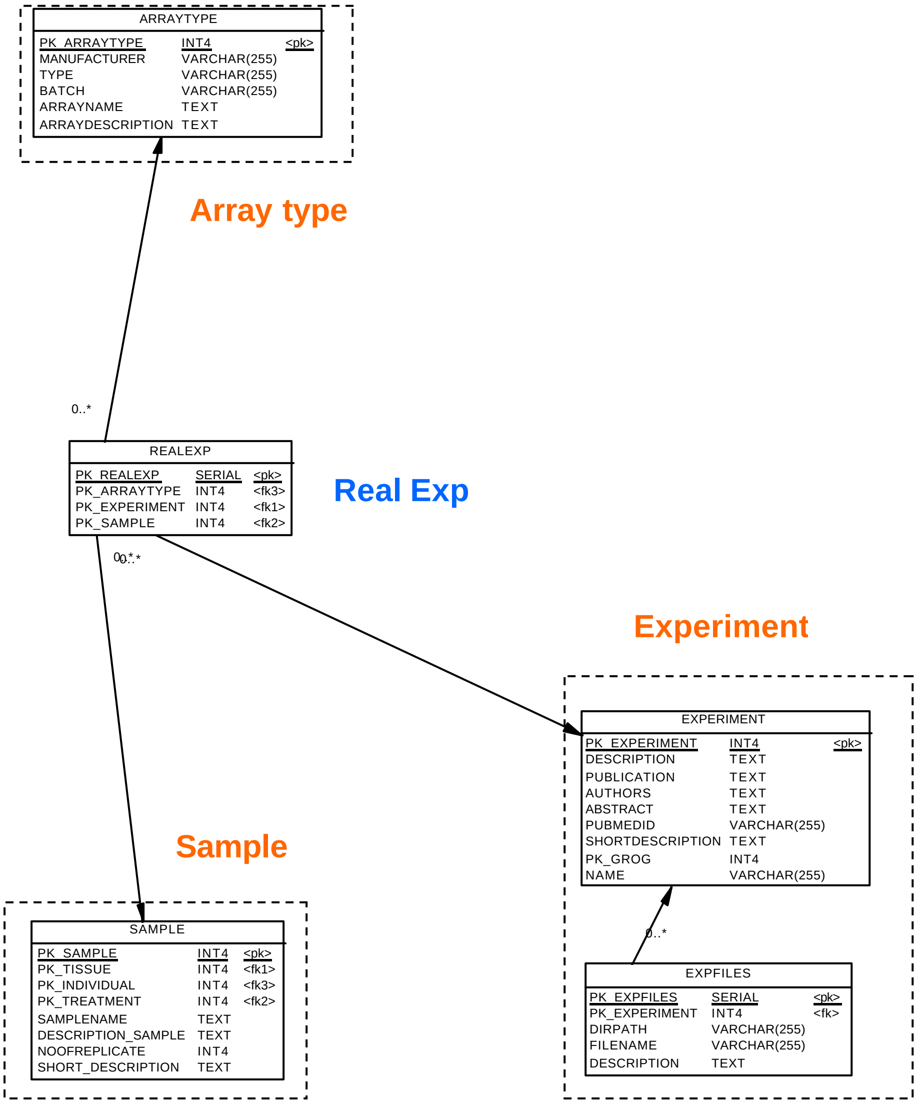


Figure 40. Schematic representation of links between Probeset, Arraytype, Organism, Sample and Experiment modules.
Three modules Arraytype, Sample and Experiment are linked two each other through a link table "Real Exp". "pk" indicates primary key, which should be unique to the table and "fk" indicate foreign key. Int, text, varchar indicates data types.

Data module schema represented in dashed line in (Figure 41) is unique to RETINOBASE, this kind of relational schema enabled us to query the gene expression and cluster data efficiently. Data module is composed of "Signal intensity", "Ratio" and "Analysis software" tables. These tables store gene expression data at two levels, one at the level of signal intensity, other at the level of fold change (detailed in next section data analysis) and also information about the analysis software used to get this data. In addition to signal intensity and fold change data, the module also contains information about calls, standard error and *p*-value for those experiments which are analyzed by the normalization software dChip (Li and Wong, 2001) and MAS5.0 (Hubbell *et al.*, 2002) respectively. The way the "Data" module is connected to "Real exp" is also unique, each value in "Signal intensity" and "Ratio" table are link to one and two "Real exp" because ratio is always compared between two samples (Figure 41). To simplify we could say that a "realexp" correspond to one hybridization process.

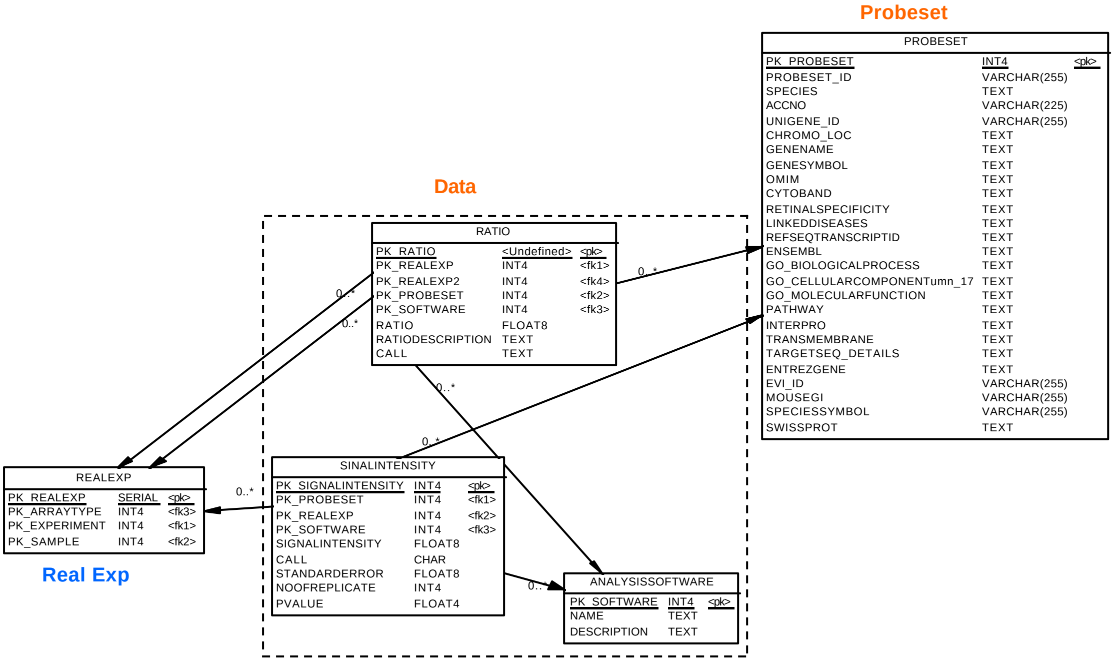


Figure 41. Schematic representation of "Data" module.
"Data" is linked to "probeset" module in order to get information about probeset and linked to "Realexp" module to get information of Sample, Array type and Experiment. "pk" indicates primary key, which should be unique to the table and "fk" indicate foreign key. Int, text, varchar indicates data types.

A cluster is a set of probesets with similar or specific profiles within one experiment analyzed by one normalizing method and one clustering program. The **Cluster** table links an experiment, probeset, analysis software and cluster software tables. "Cluster software" table has information about various algorithms used for clustering gene expression data (Figure 42). This module in RETINOBASE allows us to do meta-analysis of all clusters in all experiments analyzed by different methods and clustered using different algorithms.

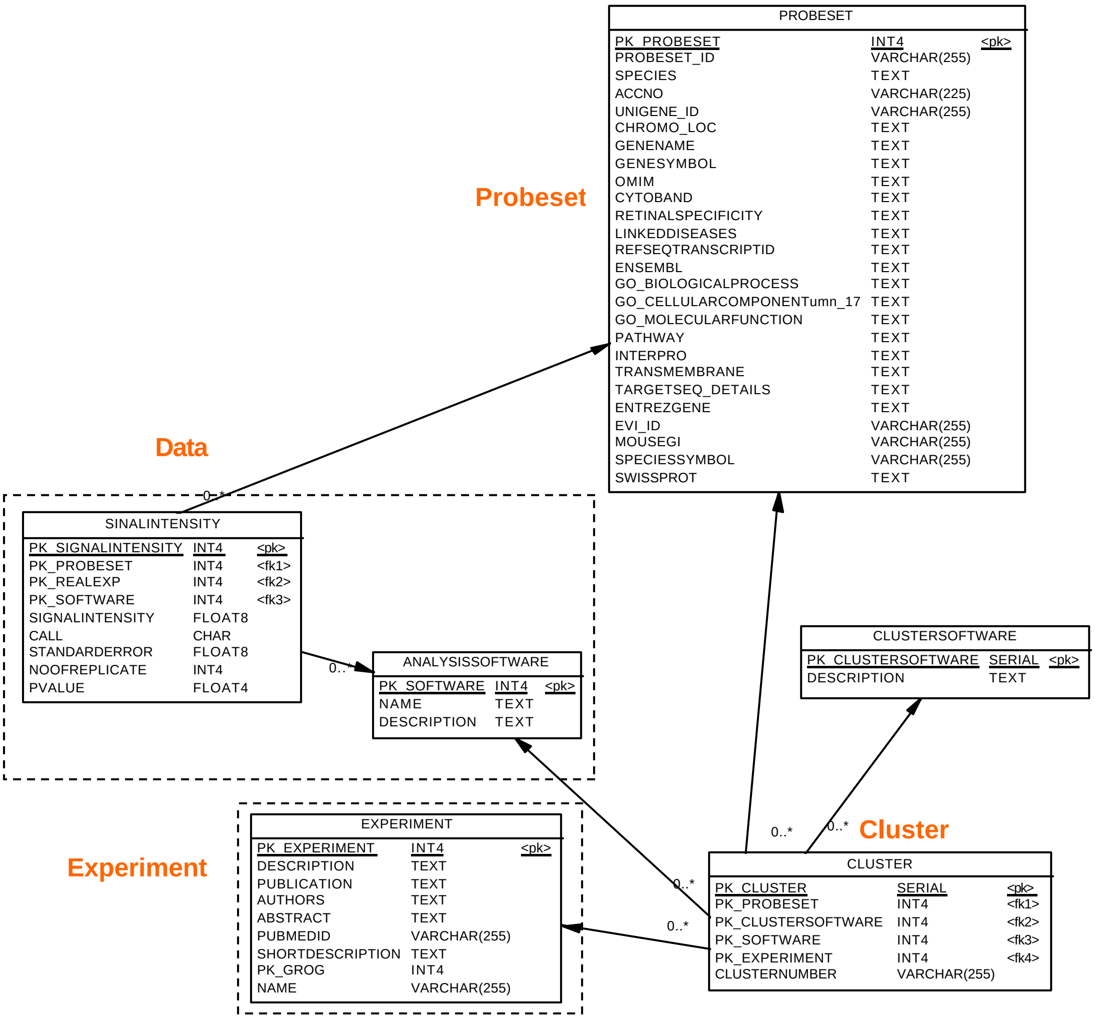


Figure 42. Schematic representation of "Cluster" module.
"Cluster" is linked to "probeset" module in order to get information about probeset, linked to "Data" module to get information about data analysis software and link to experiment helps in identify to which experiment this cluster belong. "pk" indicates primary key, which should be unique to the table and "fk" indicate foreign key. Int, text, varchar indicates data types.

.

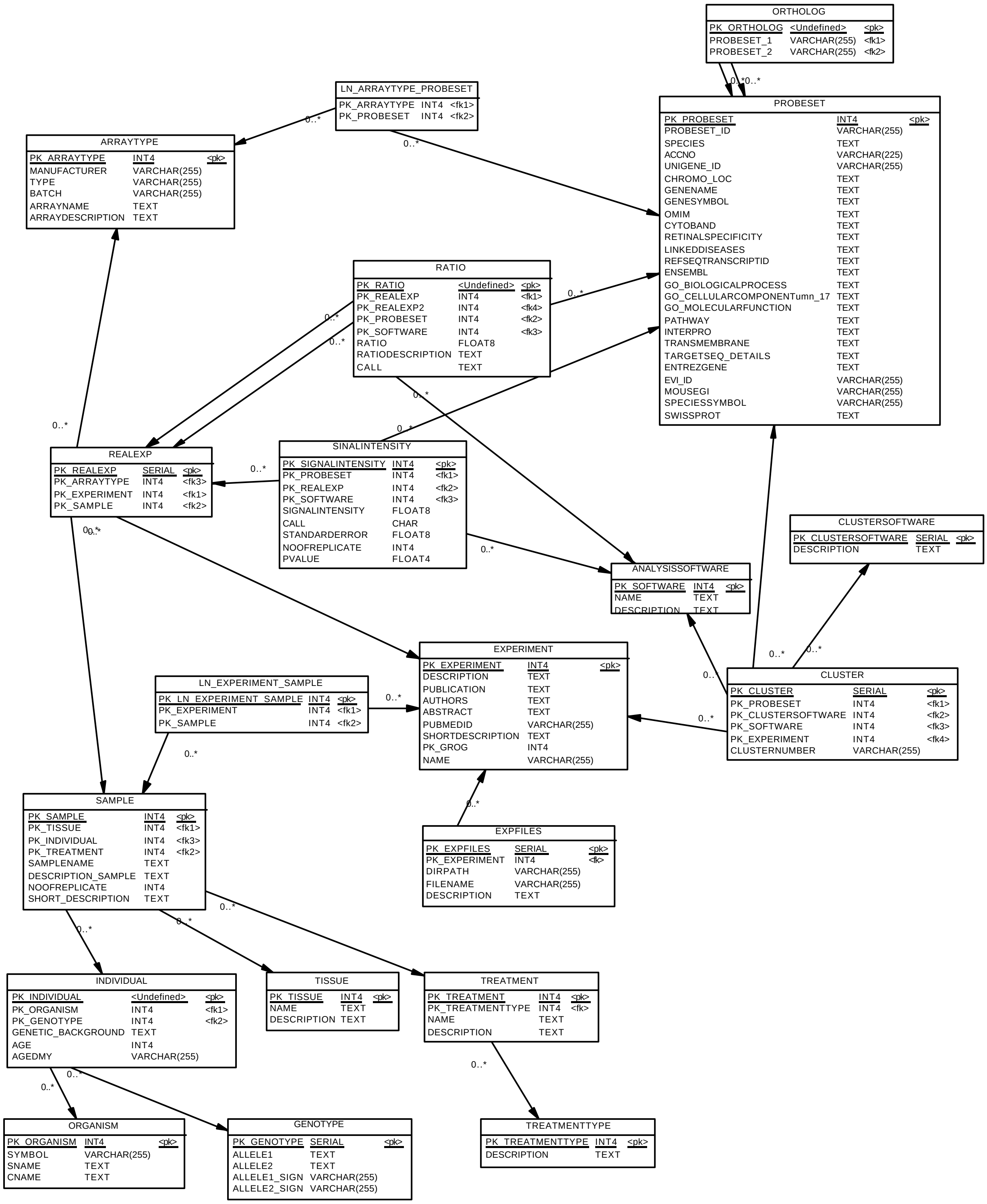


Figure 43. Overall relational schema of RETINOBASE.

10.4 Experiments available in RETINOBASE

At present, RETINOBASE has 30 experiments from 6 different organisms (*Drosophila*, *Danio rerio*, *Rattus norvegicus*, *Mus musculus*, *Gallus gallus* and *Homo sapiens*) of which 21 (detailed in publication 3) are public and downloaded from GEO (Barrett et al., 2007) and 9 are proprietary (2 from Dr. Peter Humphries's laboratory and 7 from Dr. Thierry Leveillard's laboratory). Of the 30 experiments, a majority (21) are from *Mus musculus*, 3 each from *Rattus norvegicus* and *Homo sapiens*, 2 from *Danio rerio* and 1 each from *Drosophila* and *Gallus gallus*. Due to the difficulty in obtaining samples, just 3 experiments from human origin are available, among which 2 are from cell lines and 1 from patient's sample. This emphasizes that, like numerous other human diseases, retinal and eye diseases must be studied through animal models mimicking the progression and major injuries observed in human. In this context, RETINOBASE contains many data from mouse models, with 7 knock out mice and involving well studied retinal genes (*rho-/-*, *rd1-/-*, *rpe65-/-*, *rd7-/-*, *Nrl-/-* and *trbeta-/-*) and one natural model for glaucoma (DBA/2J). In addition, a transgenic mouse model (myocilin which might provide a possible role of myocilin in the pathogenesis on glaucoma). The zebrafish model was used to study the RPE (Retinal Pigment Epithelium) markers. Drosophila was used to study eye development pathway.

Considering the type of experiments, about 10 time-series experiments are available allowing precise analyses of the establishment and progression of a disease from embryonic or birth day up to the adult. In addition, 4 experiments concerning treatment with different type of factors such as neurotrophic factors and Transforming growth factor beta are available.

RETINOBASE contains more diverse sets of experiments performed on variety of model systems than other specialized databases such as CGED (Cancer Gene Expression Database) (Kato *et al.*, 2005)or SIEGE (Smoking Induced Epithelial Gene Expression) (Shah *et al.*, 2005), since these address specific problems related to one particular species.

10.5 Data analysis of the experiments

Datasets were obtained at two levels, one at the level of the "raw data" embedded in CEL files (with the measured intensities, locations on array hybridized…) that can be further analyzed by different normalizing software and the other, at the level of signal intensities that cannot be further analyzed by the normalization software (see Chapter 4 in introduction). All proprietary and public data are obtained as CEL files and of the 21 public experiments, one experiment has partial data. According to our main objectives of designing a database not only dedicated to storage and querying but also to cross-comparison of data treatment strategies, the data obtained as CEL files were pre-processed using up to 3 different normalization programs [RMA (Irizarry *et al.*, 2003b), dChip (Li and Wong, 2001) and MAS5 (Hubbell *et al.*, 2002)]. We chose to exclude from RETINOBASE, the results, treatments, list of deregulated genes from the published experiments, as these data frequently result from data processing involving various levels of human expertise that are hard to compare to automated analyses. Nevertheless the link to the original articles are provided.

The normalization procedures are performed outside RETINOBASE using the R statistical package (http://www.r-project.org) and Bioconductor (Gentleman *et al.*, 2004) and using the Report Generator (which allows to run routine statistical analysis using R via predefined analysis scenarios in a local and independent manner) (Raffelsberger *et al.*, 2007) developed, in the laboratory, by Wolfgang Raffelsberger. The PDF file containing the results (text, figure, tables) from Report Generator for a particular experiment is stored in RETINOBASE. Similarly pre-processing quality control reports are generated using affyQCReport (an R package that generates quality control reports for Affymetrix array data) (Gautier *et al.*, 2004; Wilson and Miller, 2005).

After preprocessing, the resulting signal intensities are uploaded to RETINOBASE using SQL (Structured query language) scripts via pgAdmin III (pgAdmin III is the most popular and feature rich Open Source administration and development platform for PostgreSQL). The fold-change in gene expression is calculated as the ratio between the signal intensities of a given gene in the treated (or knockout) model and the control. In the case of experiments performed in replicate, signal intensities are averaged before calculation of the ratios and finally incorporated into RETINOBASE.

All the experiments in RETINOBASE are clustered using three different methods: (i) K-means/TMEV, (ii) K-means/FASABI and (iii) Mixture models/FASABI. The K-means/TMEV method is used in the free, open-source system for microarray data management and analysis TMEV (Saeed *et al.*, 2006). The two other methods, K-means/FASABI and the Mixture models/FASABI method (McLachlan and Basford,

1988) implemented in the in-house FASABI (Functional And Statistical Analysis of Biological Data) software developed by Adeline Legrand and Nicolas Wicker. The difference between the two K-means methods is that K-means/TMEV is based on dot product (Soukas *et al.*, 2000) whereas K-means/FASABI is based on density of points clustering (Wicker *et al.*, 2002). In both K-means methods we can define the number of clusters we want, whereas the mixture model method algorithm generates it automatically.

All the different treatments implemented in RETINOBASE implies that a given gene might result in a complex set of data related not only to the distinct probesets associated to the genes but also to the different normalization and clustering methods (Figure 44).

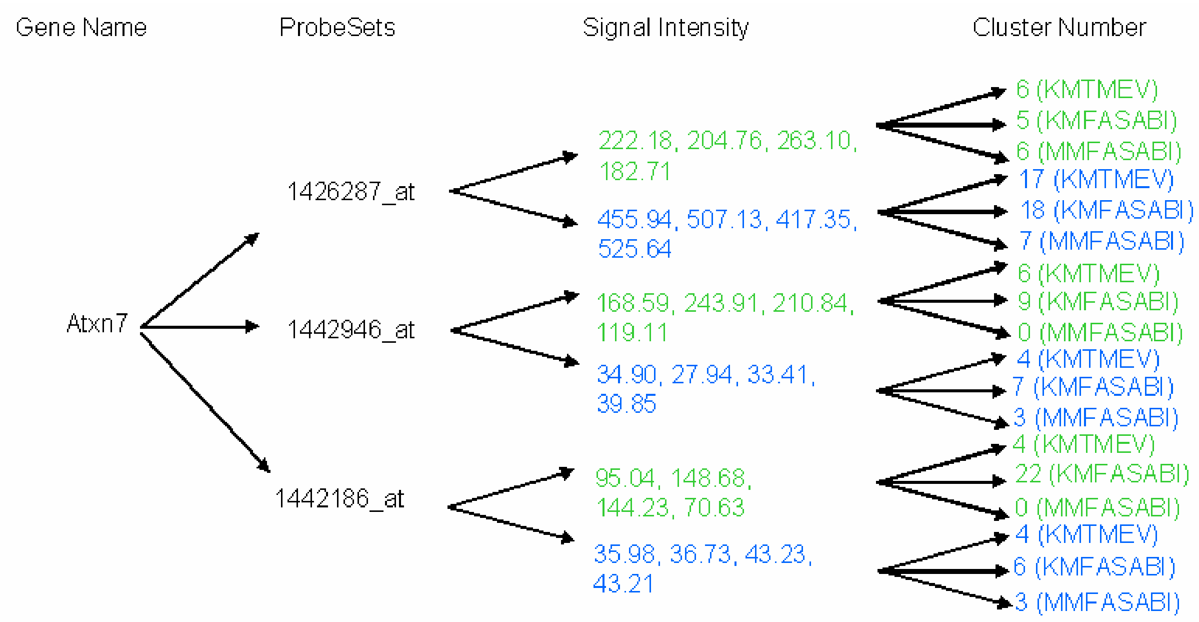


Figure 44. Different types data of associated with single gene *Atxn7* in one experiment and one sample

Experiment "Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors". Array type used in this experiment is Mouse Genome 430 2.0 array. Sample obtained from untreated Postnatal day 10 C57BL/6(WT)-Gfp old mouse photoreceptors. Blue and green colour indicates data from RMA and dChip normalization methods respectively. KMTMEV, KMFASABI and MMFASABI is K-means clustering method from TMEV, K-means clustering from FASABI and mixture model clustering method from FASABI respectively.

10.6 Querying the RETINOBASE

As described in publication 3, the six major entry points in the RETINOBASE querying system are: Experiment, Probeset or Gene, Array type, Signal intensity, Ratio and Cluster.

The Querying system is based on two types, simple or combined queries. Simple queries are performed on one table, whereas the combined queries are based on many tables.

Simple Queries

All information associated with experiment, gene or probeset and array type can be directly queried through their corresponding tables.

Examples:

- ➢ Experiment details
- ➢ Gene information of given probeset
- ➢ Description of array types.

Combined Queries

The relational database allows us to run cross queries between all the tables. For example, to display all the signal intensities of a given gene in a given experiment according to one analysis software, a combined query uses information from the probeset, sample, experiment, signal intensity and analysis software tables.

10.6.1 Gene Information

To access gene information, we offer three different query options - “Gene Query”, “Ortholog Query” and “Blast Query”. “Gene Query” and “Ortholog Query” accept the gene name, symbol, Affymetrix Probe Set ID, Refseq or Unigene IDs as input, whereas “Blast Query” accepts one sequence in FASTA format. At present “Blast Query” accepts only one sequence and the search is performed either in RefSeq or EVI-Genoret specific blast database.

In GeneChip arrays, there are many probesets that are not well annotated, in the case where there is a gene name or symbol available for the probeset, the user can alternatively enter Refseq or Unigene IDs in order to access the gene expression information.

"Ortholog Query" is useful in cross-referencing probe sets between different Affymetrix Genechip arrays. This query was made possible by linking tables "probeset" and "Ortholog".

Using the above queries one can access information such as chromosomal location, linked retinal diseases, cellular localization and gene ontologies for a given gene. Furthermore, gene details returned from these queries are linked to external databases such as (Figure 46):

1) GeneCards (an integrated database of human genes that includes automatically-mined genomic, proteomic and transcriptomic information, as well as orthologies, disease relationships, SNPs, gene expression, gene function) (http://www.genecards.org/index.shtml)

2) Entrez gene is a searchable database of genes, from RefSeq genomes, and defined by sequence and/or located in the NCBI Map Viewer (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene).

3) ADAPT mapping viewer that is particularly interesting since it describes the many-to-many relationships between Affymetrix probesets transcripts and genes, by directly mapping every probe against publicly available mRNAs/cDNA sequences from RefSeq and Ensembl (Leong *et al.*, 2005) (Figure 45).

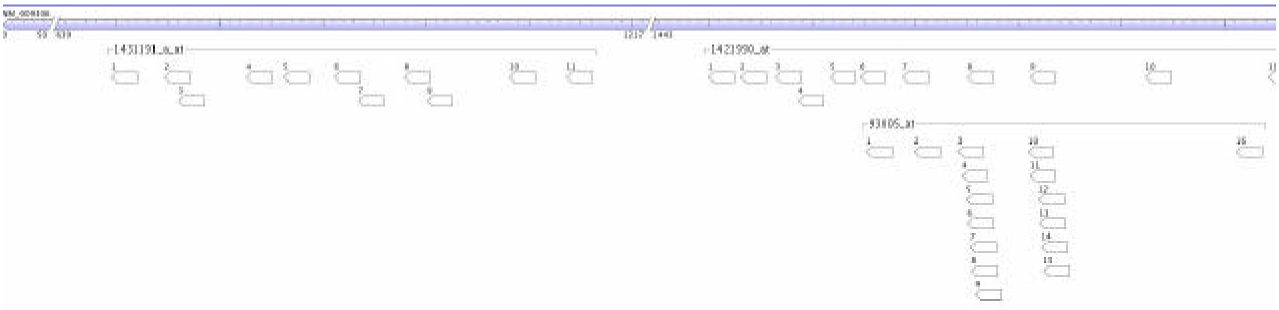
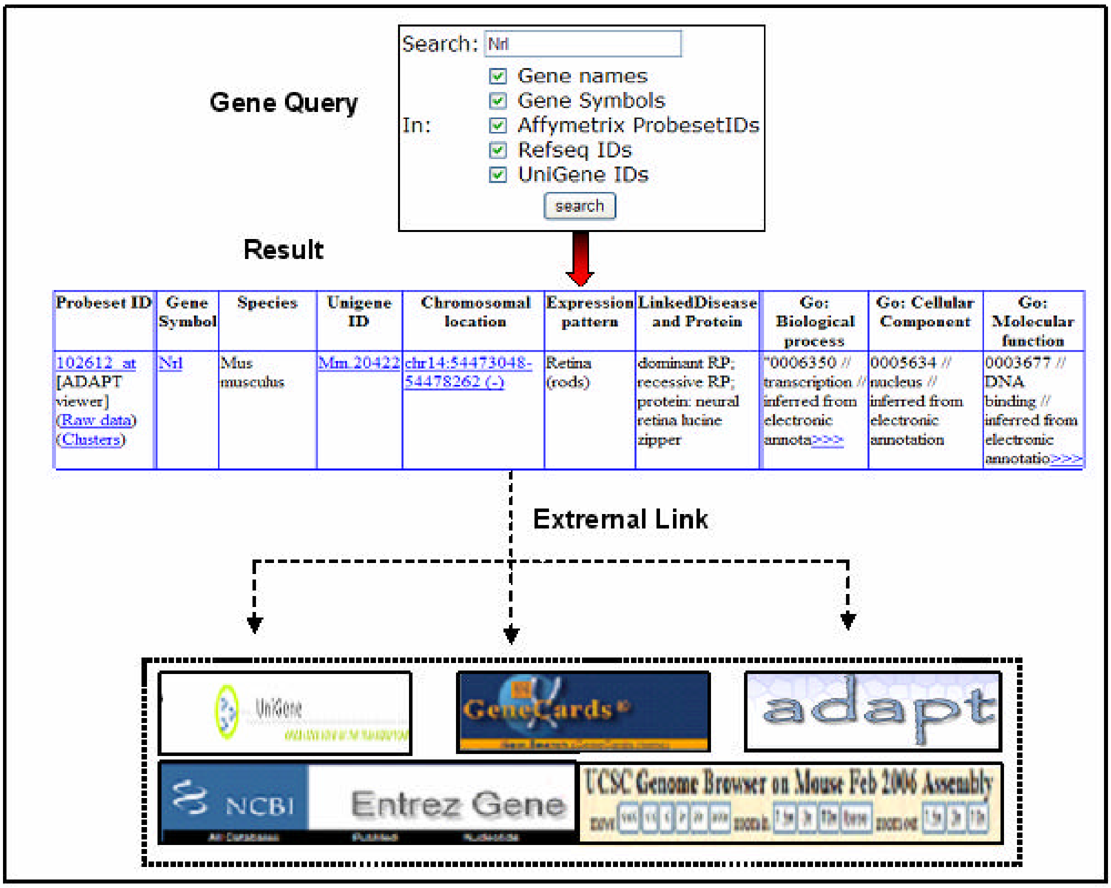


Figure 45. Visualization of individual probes of a probeset through ADAPT mapping viewer. The blue bar on top (with arrow head pointing left) represents a reverse strand transcript and the small empty arrows represent the probes spanning the entire sequence.

4) UniGene (provides information for a set of transcript sequences that appear to come from the same transcription locus (gene or expressed pseudogene), together with information on protein similarities, gene expression, and genomic location) (http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene)

5) UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgGateway) that yields more information on gene or probeset localization on genome (Figure 46).



Gene Query

Search: Nrl

In:
☑ Gene names
☑ Gene Symbols
☑ Affymetrix ProbesetIDs
☑ Refseq IDs
☑ UniGene IDs

search

Result

Probeset ID	Gene Symbol	Species	Unigene ID	Chromosomal location	Expression pattern	LinkedDisease and Protein	Go: Biological process	Go: Cellular Component	Go: Molecular function
102612_at [ADAPT viewer] (Raw data) (Clusters)	Nrl	Mus musculus	Mm.20422	chr14:54473048-54478262 (-)	Retina (rods)	dominant RP; recessive RP; protein: neural retina lucine zipper	"0006350 // transcription // inferred from electronic annota>>>	0005634 // nucleus // inferred from electronic annotation	0003677 // DNA binding // inferred from electronic annotatio>>>

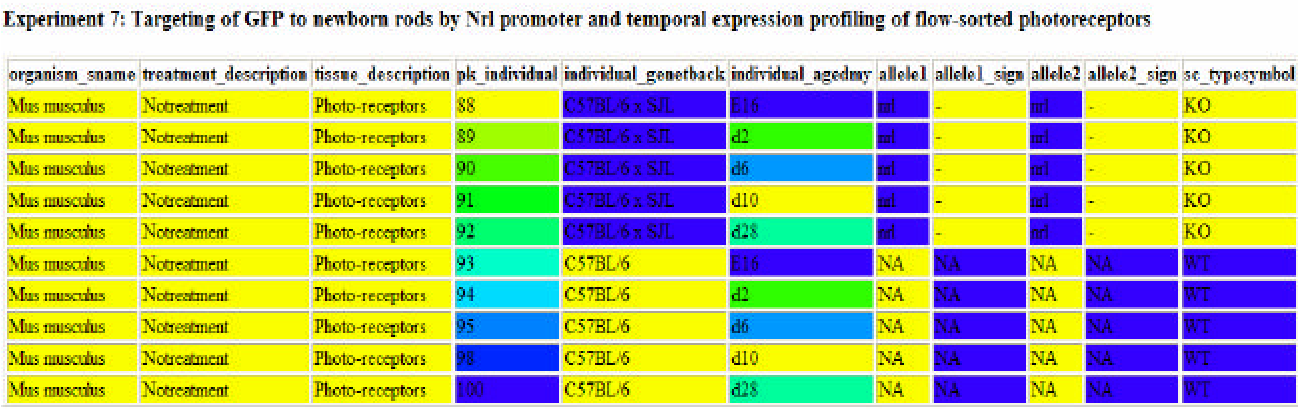
Figure 46. Gene information Query
"Gene Query" yields information such as Unigene ID, chromosomal location, Entrez gene, expression pattern, linked diseases and gene ontology. Dotted line indicates links to external databases.

Further through the "Gene Query" module, RETINOBASE allows the user to access the signal intensities and cluster information for a given gene (cluster number, software used for clustering and information about other genes present in the same cluster) for all experiments. Further the same information (signal intensity, cluster and fold change) can also be accessed from "Signal intensity or Cluster query" and "Fold change" querying systems.

RETINOBASE also provides information regarding expression of approximately 200 retinal genes specific to certain types of cell, such as photoreceptors, Muller cells or retinal sphere cells. This information is obtained from the literature, currently there is no automatic process for automatic updating of this information into the database.

10.6.2 Experiment information

User has access to a synthetic view of an experiment and its associated information about sample, organism, tissue, etc, through experiment details option (Figure 47).



Experiment 7: Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors

organism_sname	treatment_description	tissue_description	pk_individual	individual_genetback	individual_agedmy	allele1	allele1_sign	allele2	allele2_sign	sc_typesymbol
Mus musculus	Notreatment	Photo-receptors	88	C57BL/6 x SJL	E16	nrl	-	nrl	-	KO
Mus musculus	Notreatment	Photo-receptors	89	C57BL/6 x SJL	d2	nrl	-	nrl	-	KO
Mus musculus	Notreatment	Photo-receptors	90	C57BL/6 x SJL	d6	nrl	-	nrl	-	KO
Mus musculus	Notreatment	Photo-receptors	91	C57BL/6 x SJL	d10	nrl	-	nrl	-	KO
Mus musculus	Notreatment	Photo-receptors	92	C57BL/6 x SJL	d28	nrl	-	nrl	-	KO
Mus musculus	Notreatment	Photo-receptors	93	C57BL/6	E16	NA	NA	NA	NA	WT
Mus musculus	Notreatment	Photo-receptors	94	C57BL/6	d2	NA	NA	NA	NA	WT
Mus musculus	Notreatment	Photo-receptors	95	C57BL/6	d6	NA	NA	NA	NA	WT
Mus musculus	Notreatment	Photo-receptors	98	C57BL/6	d10	NA	NA	NA	NA	WT
Mus musculus	Notreatment	Photo-receptors	100	C57BL/6	d28	NA	NA	NA	NA	WT

Figure 47. Synthetic view of experiment 7: "Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors". Organism is *Mus musculus*, tissue is photo receptors, there are two different genetic backgrounds, 5 different days and comparison is between Nrl(-/-) against wildtype. Similar post natal days for Nrl (-/-) and the wildtype have same colours e.g. Embryonic day 16 (E16) is coloured in dark blue, postnatal day 2 in green.

10.6.3 Signal intensity system analysis

This module has two different ways to visualize signal intensity data "Signal intensity system analysis" (extract data in tables) and "Data visualization" which is both a query and visualization option (extract data in graphical format) (this option is detailed in further section). "Signal intensity Query" provides gene expression information at the level of signal intensities for single or multiple genes in one or more experiments. "Cluster Query" is unique to RETINOBASE and provides information of similar expression patterns of related genes across varied conditions and genetic backgrounds. It also identifies any two given genes in the same cluster in one or more experiments.

Overview of querying speed of RETINOBASE: Currently, RETINOBASE contains approximately 30 million gene expression values resulting from 540 hybridizations and 242 samples. We tested the querying speed of the "Signal intensity Query" by using the Leucine rich repeat containing G protein coupled receptor 5 (*Lgr5*) gene. This gene has 3 probesets on GeneChip (MG 430 2.0). To extract data at the level of signal intensities using one of the three probeset ids as input in an experiment with 27 samples with 4 replicates each, it took 3516 milliseconds (ms) and it retrieved 132 rows. Similarly when the input was gene symbol, it took 10360 ms and retrieved more rows for the same query.

10.6.4 Fold change System Analysis

Gene expression information at the level of fold change is provided for single or multiple genes in one or more experiments. We provide a Fold change query option to the user with two different ways to query gene expression data 1) the user enters either gene name, symbol or probeset and then enters the experiment in order to visualize the fold change of that particular gene. 2) Alternatively he can choose one experiment or many experiments and specify the criteria (greater and/or less given fold change) in order to visualize all the genes that meet this criteria.

To extract data at the level of fold change using probeset id as input in an experiment with 27 samples, each “Fold change” query took 688 ms and retrieved 26 rows. A similar query using gene symbol as input took 3891 ms and retrieved more rows. When we used a generalised query to retrieve all genes that have fold change greater than 15 in the same experiment as above, the query time is 4547 ms.

10.7 Signal Intensity visualization system

To offer an ergonomic graphical representation of the complex data associated with an experiment, we developed a useful visualization tool, which runs cross queries with all important tables in our database (Figure 48).

This web application uses the AJAX (Asynchronous JavaScript and XML) web development technique to increase the interactivity and functionality of the page.

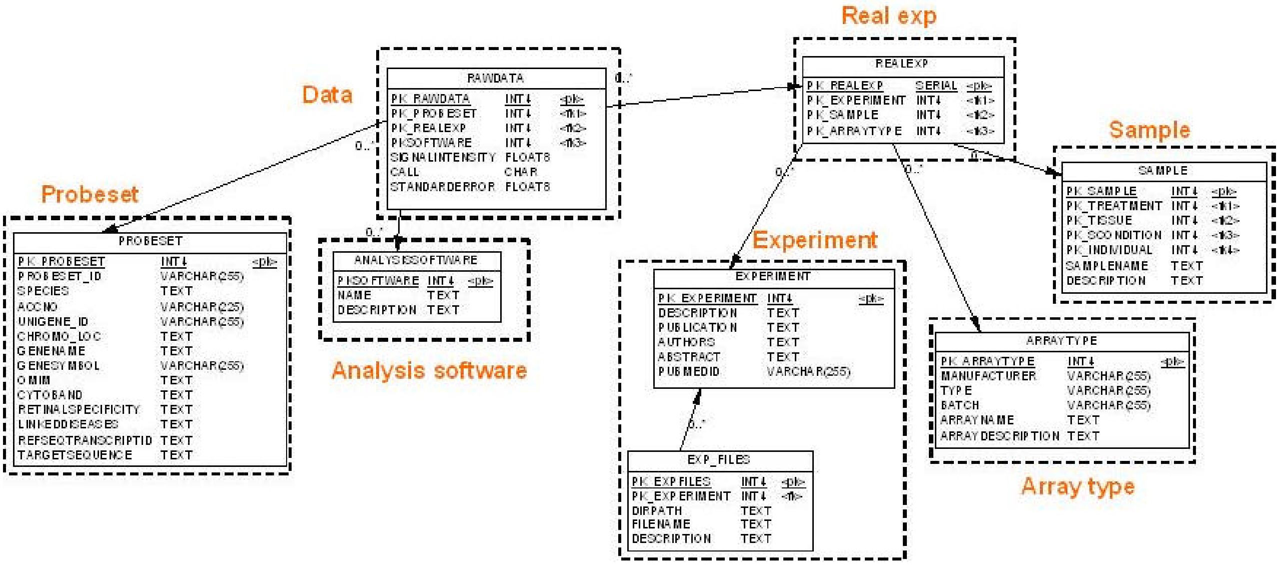


Figure 48. Relational schema under data visualization module.

This visualization module not only presents background corrected and normalized data graphically but also provides quality of replicates in an experiment by providing Coefficient of Variation (CV) between replicates. The CV is calculated on the fly using the formula CV = SD/X, where SD is standard deviation between replicates and X, the mean between the replicates.

Data visualization options are provided in a step-by-step query method (Figure 49). First the user can choose the Experiment, thus the array type is automatically provided. After selecting the array type, all the samples in the given experiments are detailed according to the name and order introduced in RETINOBASE. Another option is to choose an array type and RETINOBASE provides all the experiments related to this array type allowing the user to choose any one of the experiments. It is important to note that after experiment and array type choice, the user can rename or sort the distinct samples according to his convenience. Then, the user can enter the gene of interest, in order to visualize the quality control information. Finally, the user can choose to visualize the data at signal intensity level as "radar plots"(Figure 51) or "histograms" (Figure 52) or "standard deviation" graphs (Figure 53). The three different graphs are generated on the fly. We chose to include three visualization options as each of them would help a user gain insight into various aspects of the results obtained. In general, results from a time series experiment are best visualized as radar plots and also provide an opportunity to quickly compare the data. Whereas data obtained from an experiment with only two conditions (control and test) are better visualized as histograms. Standard deviation plots, on the other hand, allow users to asses the quality of the replicates in the experiment.

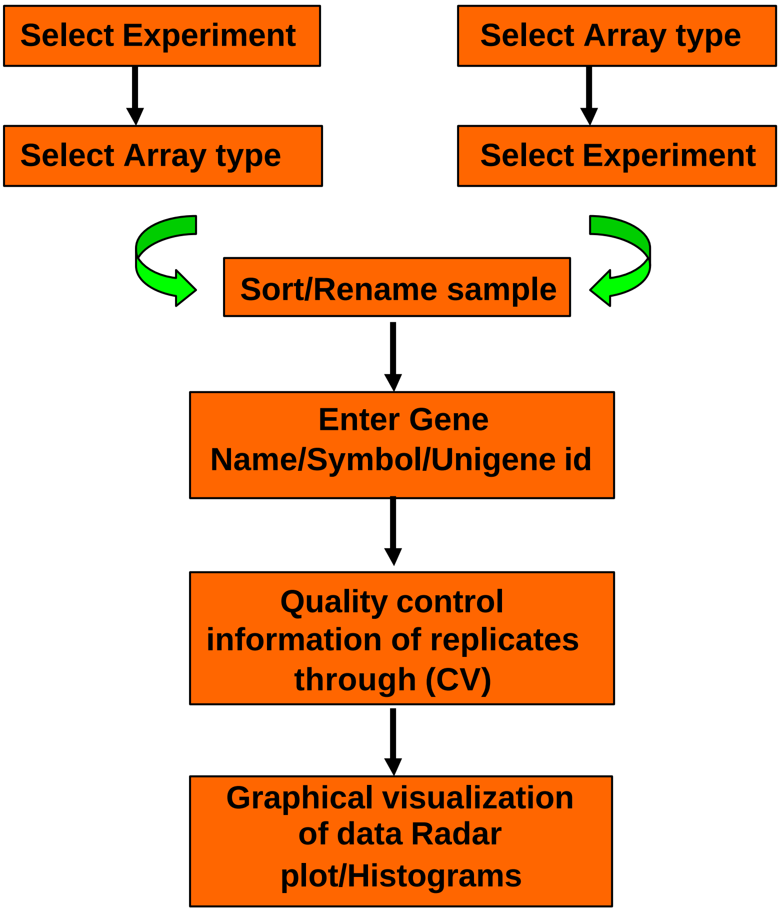

Select Experiment
Select Array type
Select Array type
Select Experiment
Sort/Rename sample
Enter Gene Name/Symbol/Unigene id
Quality control information of replicates through (CV)
Graphical visualization of data Radar plot/Histograms

Figure 49. Step-by-step procedure of data visualization query
The user can choose any of the available experiments, array type. Then, all the samples available are shown and can be sorted or renamed. After this step, the user can enter the "Gene name" or "Unigene id" in order to visualize its expression profiles. Quality information about the gene in replicate samples and in the chosen experiment is provided via Coefficient of Variation (CV).

To illustrate that different probesets of the same gene can show different signal intensities in one experiment analyzed by different normalization methods, in this case RMA (Figure 50) and dChip (Figure 51), we chose Synaptotagamin1 (*syt1*) gene and experiment "Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors". This experiment was performed on Mouse Genome 430 2.0 array. In this array there are six different probesets corresponding to same *sty1* gene. The difference in signal intensity observed may be primarily due to hybridization problems, to position of the probesets from 3' end of the gene or may represent different splice variants of the same transcript.

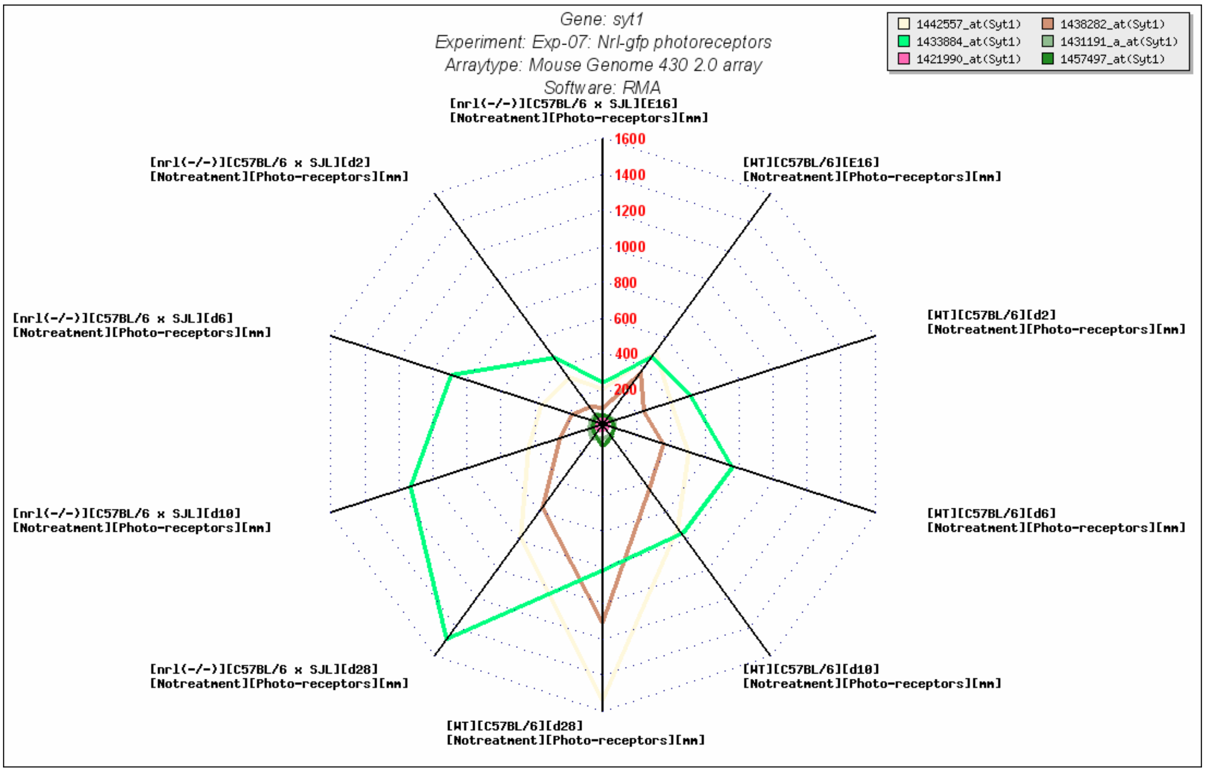


Figure 50. Signal intensity visualization result as radar plot.
Expression profile of 6 Probe Sets of Synaptotagamin1(*syt1*) gene in the nrl (-/-) and wildtype (wt) at various time points (Embryonic day 16, post natal day 2, 6, 10 and 28) in 6 different colors using RMA method. Left side represents nrl -/- knockout and right side wild type.

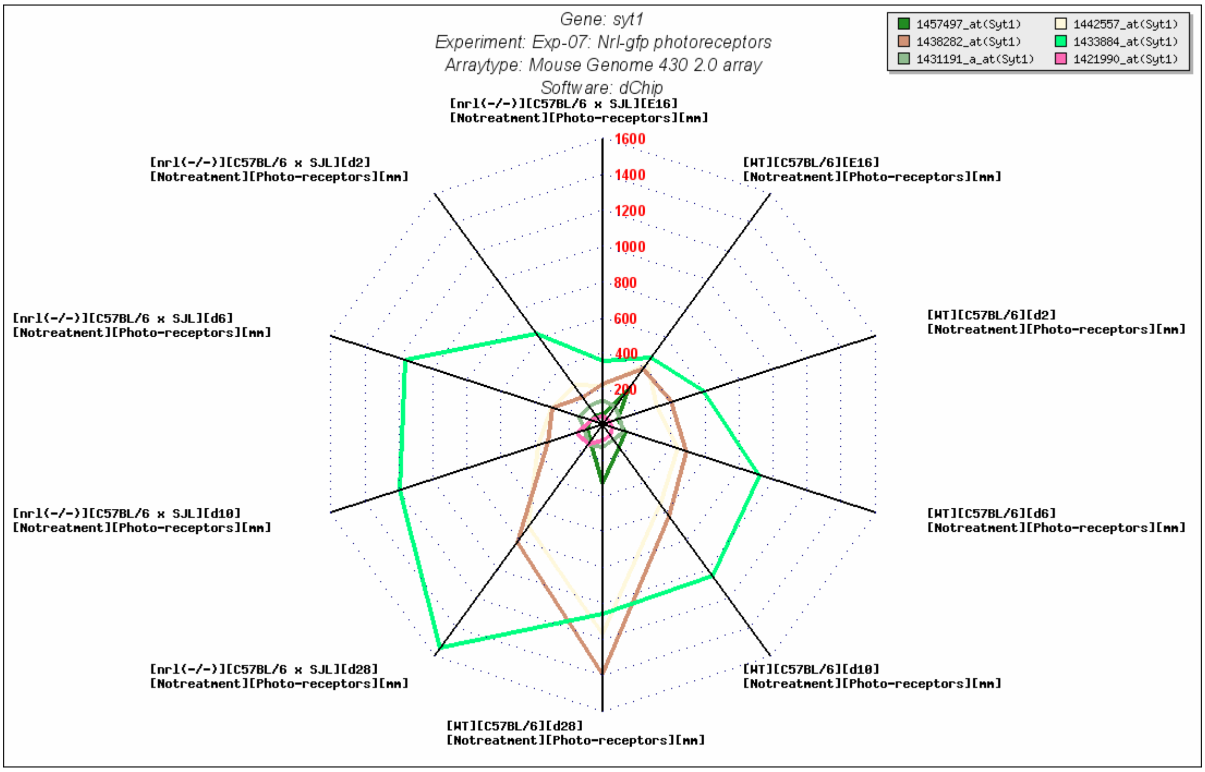


Figure 51. Signal intensity visualization result as radar plot.
Expression profile of 6 Probe Sets of Synaptotagamin1(*syt1*) gene in the nrl (-/-) and wildtype (wt) at various time points (Embryonic day 16, post natal day 2, 6, 10 and 28) in 6 different colors using dChip method. Left side represents nrl -/- knockout and right side wild type.

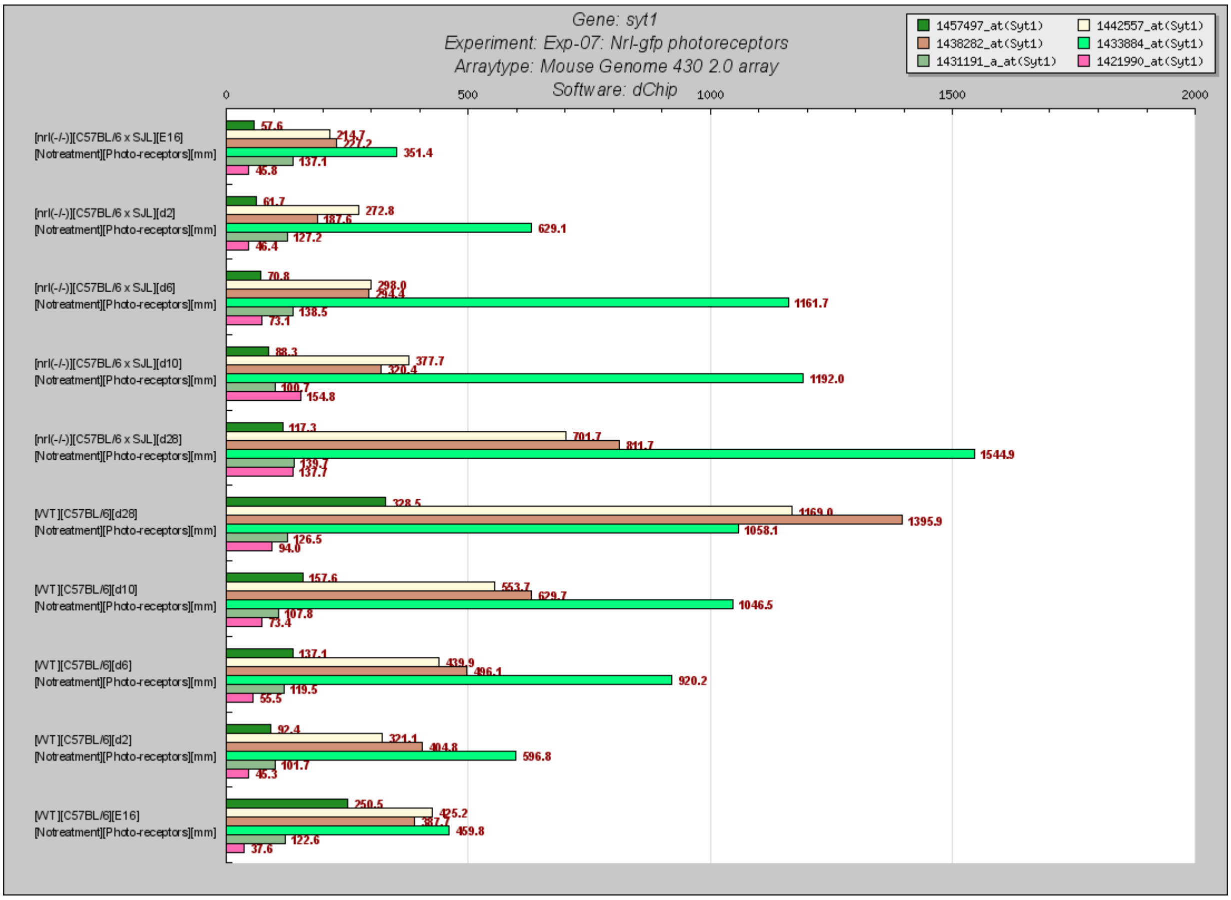


Figure 52. Signal intensity visualization result as histograms.
Expression profile of 6 Probe Sets of Synaptotagamin1(*syt1*) gene in the nrl (-/-) and wildtype (wt) at various time points (Embryonic day 16, post natal day 2, 6, 10 and 28) in 6 different colors using dChip method.

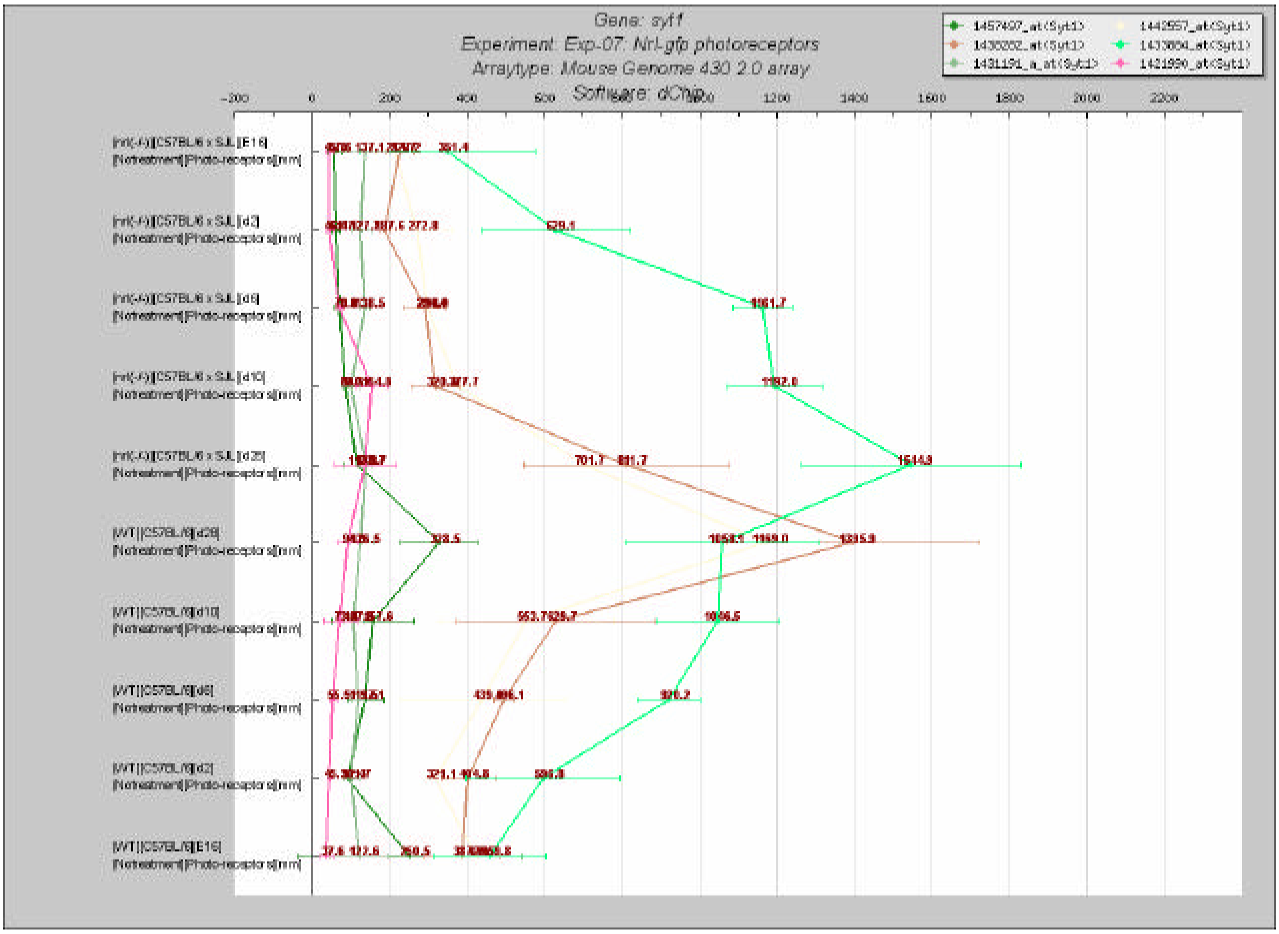


Figure 53. Signal intensity visualization result as standard deviation graph. Expression profile of 6 Probe Sets of Synaptotagamin1(*syt1*) gene in the nrl (-/-) and wildtype (wt) at various time points (Embryonic day 16, post natal day 2, 6, 10 and 28) in 6 different colors using dChip method.

10.8 Downloading results and User manual

In order to allow the users to further compare and interpret data, we provided an option to download results from all querying modules available in RETINOBASE in Comma Separated Value (.CSV) file format using the "Download results" option. At present we do not provide options to download data at the level of CEL because they can be downloaded from the public repositories.

We created a user manual that provides a detailed description of the utilities and functionalities available in RETINOBASE and is available at http://alnitak.u-strasbg.fr/RetinoBase/usermanual_retinobase.doc.

10.9 Future developments

RETINOBASE is under constant development, including addition of new experiments when available. In addition, data from proprietary experiments can be accessed on approval by individual researchers and will be made generally available after publication. In order to protect each experiment and its associated data, we have introduced a login and password system allowing the management of access rights and visibility groups. This 'people management' system was initially provided by the Genoret Database. RETINOBASE now has its own people management system and a coupled sharing of users still permits the Genoret users to access the RETINOBASE native access rights system. Several functional enhancements are also planned for the future. We will continue to refine and update RETINOBASE with respect to data retrieval, mining and visualization options. At present, there is no direct link from BLAST queries to the gene expression data in RETINOBASE, and in our future developments we plan to provide this option.

Meta-analyses using the MDF method will be performed on retinal data, that will subsequently be incorporated into RETINOBASE. We also further plan to link various queries available in RETINOBASE to query this meta-analyzed data. In order to increase the interactivity and functionality of the various query systems, we plan to develop step-by-step querying options as in the case of data visualization option. This procedure also helps in faster retrieval of data.

10.10 Conclusions

A better understanding of normal tissue functions and also how dysfunction leads to a pathological condition could be achieved by identifying gene expression patterns. Our original goal was to design RETINOBASE as a dedicated storage and analysis platform for automated and systematic methods for analyses and organization of gene expression data. The lack of a microarray database dedicated to the retina led us to build RETINOBASE which stores retinal gene expression data. Datasets in RETINOBASE were analyzed using various normalization methods and further clustered using a number of different algorithms. This enabled RETINOBASE to serve as a comparison platform for these different methods.

Our database, with different types of query options and powerful visualization tools, allows comprehensive analysis of biological mechanisms/pathways of the retina in normal and under diseased conditions. Today, three years since its initiation, RETINOBASE has grown to a sizeable platform that can be used to analyze, visualize and compare retinal-related data that could provide insights into retinal gene expression and related diseases.

The usefulness of the system should be further enhanced as new experiments are integrated. This is done on a regular basis, when data is made available and will continue to be done here in the lab in unison with EVI GENORET under FP7.

10.11 Example applications of RETINOBASE

The categorization of genes and their expression values obtained under various conditions in a systematic approach (from statistical analysis to storage in database) has a wide range of applications, such as understanding molecular mechanisms involved in various development stages and in disease conditions, in studies of the effects of various treatments on tissues or cells, or in developing tissue specific microarray chips.

10.11.1 RetChip

RETINOBASE has been extensively used in developing RetChip, an oligonucleotide microarray to study mouse retinal development and degeneration. RetChip contains the 1500 genes that are most relevant to retinal biology.

We started with a preliminary list of ~3800 genes comprising ~ 298 compulsory genes (which represent "true retinal" genes or genes that are absolutely necessary to the study and analysis of the expression behaviour of the retinal tissue), plus a list of genes provided by members of the RETNET and EVI-GENORET consortia, representing different families of genes (G-protein coupled receptor, Caspases, Neurotrophic factors etc) and genes identified by various transcriptomics projects.

To obtain the final list of ~1500 genes for the RetChip from this 3800 genes, we have established a filtration protocol which represents a prototypal "retinal propensity score". This protocol is partly based on the combination of 298 "Compulsory genes" and on various other methods such as the Retscope protocol, genomic data, transcriptomics analysis etc.

We based our strategy on in depth analysis of the clustering patterns and expression behaviour of the 4 transcriptomics experiments available and automatically analysed in RETINOBASE. These 4 experiments were analysed using the dChip or RMA normalization methods and clustered using 3 different clustering algorithms, as mentioned earlier in this chapter. This led to 12 combinations for these 4 experiments. From these 12 combinations, we removed clusters with cluster size ~>2000 from the analysis, since these clusters usually contain noisy data and genes with smaller signal intensities. We further went on to determine the genes which were frequently observed with a similar cluster distribution as that of the compulsory genes.

We have defined optimized criteria corresponding to the genes that exhibit a detectable signal intensity (signal intensity greater than 50) in at least twelve combinations (regardless of the normalisation procedure) and which share at least eight times (out of twelve distinct combinations) the same cluster as a compulsory gene. These criteria allowed us to retrieve 1089 genes. After this step we had only 1387 (1089 + 298 compulsory genes) which were further analysed and shown to be strongly enriched in retinal functions or genes as estimated by functional annotation and EST profiles. In order to obtain the total number of ~1500 genes, a similar approach was used but using less stringent selection criteria. This time genes should share the same cluster distribution as that of the compulsory genes at least 7 times out of twelve distinct combinations. After this we obtained ~ 1211 genes. To select the most pertinent ~120 genes within this pool of newly retrieved genes, we have developed additional filters that have been manually checked and validated. Genes were eliminated when:

1) No signal intensity was observed in one particular transcriptomic experiment (wild type versus rd1 (-/-)),
2) A null or very low "EST index" was observed (the "EST index" represents the percentage of ESTs corresponding to the genes, compared to the total number of observed ESTs), i.e. genes with few or no ESTs observed in any of the eye-related tissues were eliminated.
3) No corresponding cDNA was present in any of the mouse retinal cDNA databases provided by T. Léveillard.

These distinct values were used to calculate a final index to classify the pool of remaining genes, as well as the set of selected genes as a validation test. We took the genes with the best index to reach the total number of 1500 genes, which were retained for the first release of the RetChip.

Publication n^o 3

RETINOBASE: a web database, data mining and analysis platform for gene expression data on retina

Ravi Kiran Reddy Kalathur [1], Nicolas Gagniere [1], Guillaume Berthommier [1], Wolfgang Raffelsberger [1], Raymond Ripp [1], Thierry Léveillard [2], Olivier Poch[1§]

[1]Laboratoire de Biologie et Genomique Structurales, Institut de Génétique et de Biologie Moléculaire et Céllulaire, CNRS/INSERM/ULP, BP 163, 67404 Illkirch Cedex, France

[2] Inserm U592 Universite Pierre et Marie Curie, Laboratoire de Physiopathologie Céllulaire et Moléculaire de la Retine, Hopital Saint-Antoine, Paris, France

[§]Corresponding author: Olivier Poch, email: poch@titus.u-strasbg.fr

Email addresses:

RKRK: ravi@igbmc.u-strasbg.fr

NG: gagniere@igbmc.u-strasbg.fr

GB: berthomg@igbmc.u-strasbg.fr

WR: wraff@igbmc.u-strasbg.fr

RR: ripp@igbmc.u-strasbg.fr

TL: Thierry.Leveillard@st-antoine.inserm.fr

Abstract

Background

The retina is a multi-layered sensory tissue that lines the back of the eye acting at the interface of input light and visual perception. Its main function is to capture photons, and convert them into electrical impulses that travel along the optic nerve to the brain where they are turned into images. It consists of neurons, nourishing blood vessels and different cell types of which neural cells predominate. Defects in any of these cells can lead to a variety of retinal diseases including age-related macular degeneration, retinitis pigmentosa, Leber congenital amaurosis and Glaucoma. Recent progress in genomics and microarray technology provides extensive opportunities to examine alterations in retinal gene expression profiles during development and diseases. However, there is no specific database that deals with retinal gene expression profiling. In this context this we have built RETINOBASE a dedicated microarray database for retina.

Description

RETINOBASE is a microarray relational database, analysis and visualization system allowing simple yet powerful queries to retrieve information about gene expression in retina. It provides access to gene expression meta-data and offers significant insights into gene networks in retina, resulting in better hypothesis framing for biological problems that can subsequently be tested in the laboratory. **Public and proprietary** data are automatically analyzed with 3 distinct methods, RMA, dChip and MAS5 then clustered using 2 different K-means and 1 mixture models method. Thus, RETINOBASE provides a framework to compare these methods and to optimize the retinal data analysis. RETINOBASE has three different modules, “Gene Information”, “Raw Data System Analysis” and “Fold change system Analysis” that are interconnected in relational schema allowing efficient retrieval and cross comparison of data. Currently, RETINOBASE contains datasets from 28 different microarray experiments performed in 5 different model systems: drosophila, zebrafish, rat, mouse and human. The database is supported by a platform that is designed to easily integrate new functionalities and is also frequently updated.

Conclusions

The results obtained from various biological scenarios can be visualized, compared and downloaded. RETINOBASE provides efficient access to the global expression profiling of retinal genes from different organisms under various conditions and can be accessed at ***http://alnitak.u-strasbg.fr/RetinoBase/.***

Background

The retina is a thin and highly structured layer of neuronal cells that lines the back of eye. Its main function is to convert light energy into an interpretable signal for cortical cells in the brain. The retina has two components - an inner neurosensory retina and an outer retinal pigment epithelium (RPE), which together form the structural and functional basis for visual perception.

The retina consists of several cell types of which neural cells predominate. Photoreceptors, bipolar and ganglion cells are three principal neuron cell types whose activity is modulated by other groups of cells such as horizontal and amacrine cells [1]. Defects in any of the above-mentioned cell types can lead to a variety of retinal diseases including age-related macular degeneration (AMD), retinitis pigmentosa (RP), Leber congenital amaurosis (LCA) and Glaucoma. These diseases may lead to partial visual loss or complete blindness, depending on the severity.

The recent progress in genomic approaches has now lead to an increase in the number of transgenic and knockout animal models that can be used to investigate the role of specific genes in retinal function and related disorders in humans, e.g., *rd1* is a mouse model for RP [2], *Nr2e3* for the Human Enhanced S-cone syndrome (ESCS) [3], *Rds* for macular dystrophy and *RPE65*$^{-/-}$ for LCA [4]. Experimental information from the above mentioned models combined with high-throughput technologies has lead to an increase in the number of experiments related to retinal gene expression

The recent development of high-throughput technologies has resulted in an enormous volume of gene expression data. General repositories such as GEO [5] and ArrayExpress [6] operate as central data distribution centres encompassing gene expression data from different organisms and from various conditions. In contrast, resources like CGED [7], SIEGE [8] and GeneAtlas [9] are specialized databases that address specific problem; CGED addresses gene expression in various human cancer tissues, SIEGE focuses on epithelial gene expression changes induced by smoking in humans and Gene Atlas provides the expression profiles of genes in various mouse and human tissues.

In order to address specific issues related to retina and to meet the needs of retinal biologists in their analysis of gene expression data, we have developed RETINOBASE, a microarray gene expression database for retina. RETINOBASE combines simplified querying, analysis and data visualization options. The integration of gene expression data from various development stages of wild type retina and from diverse conditions and genetic backgrounds will hopefully, not only increase our understanding of the physiological mechanisms involved in normal retinal tissue but also facilitate studies gene expression patterns under diverse conditions. Furthermore, RETINOBASE provides a platform for the comparison of different analysis scenarios based on various normalization methods such as RMA [10], dChip [11], MAS5 [12] and clustering methods such as K-means [13] and mixture models method [14].

Construction and content

RETINOBASE uses open-source tools. The website is powered by an Apache web server, PHP and Javascript for dynamic web pages and a PostgreSQL object-relational open source database management system (DBMS) as the back end to store data. The RETINOBASE database schema has been developed using the same philosophy as that used to design BASE [15], with enhancements to accommodate data from different platforms and also complies to the Minimum Information About Microarray Experiment (MIAME) standard [16]. It is based on a well-designed relational schema where “realexp” acts as a central table linking expression data with an experiment, sample and array type. This kind of schema helps the system to manage data efficiently, and increase retrieval speed.

RETINOBASE is designed to store gene expression profiles from a microarray experiment. We downloaded publicly available data via FTP from Gene Expression Omnibus (GEO) for 21 experiments [17-32], GEO data (GSE 1816, 4756, 1835, 3791, 2868), and additionally 8 proprietary experiments have been incorporated that can be accessed with permission from the owner of the experiment. These experiments were performed under different conditions, including knockout models, treatments and time series experiments performed on different organisms such as drosophila, zebra fish, rat, mice and human. All experiments have complete data except for 1 experiment [19] that has partial data at the level of fold change due to unavailability of raw data (.CEL) or signal intensity data. Currently, RETINOBASE contains approximately 270 million gene expression values resulting from 509 hybridizations.

Gene information

In RETINOBASE, gene annotation information obtained from Affymetrix (www.affymetrix.com) is linked to information about genes and loci causing inherited retinal diseases, obtained from the Retinal information network (RETNET) http://www.sph.uth.tmc.edu/Retnet/. RETINOBASE also provides information obtained from literature about expression of approximately 200 retinal genes specific to certain types of cell, such as photoreceptors, Muller cells or retinal sphere cells.

Data information

Raw data was obtained in two different formats, either as .CEL files (20 experiments) or at the level of signal intensities (8 experiments). Data obtained at the level of .CEL files are first analysed with three different normalization programs - RMA [10], dChip [11] and MAS5 [12] using the R statistical package (http://www.r-project.org) and Bioconductor [33], and after preprocessing, the resulting background-corrected and normalized signal intensities are automatically uploaded to RETINOBASE using SQL scripts via pgAdminIII.

Identification of control samples in an experiment facilitated incorporation of data at the level of fold change in RETINOBASE. The fold-change in gene expression is calculated as the ratio between the signal intensities of a given gene in the treated (or knockout) model and the control. In the case of experiments performed in replicate, signal intensities are averaged before calculation of the ratios. All the experiments in RETINOBASE are clustered using 3 independent methods: (i) the density of points clustering method [34] implemented in the in-house FASABI (Functional And Statistical Analysis of Biological Data) software, (ii) the dot product K-means method [35] used in TMEV, a free, open-source system for microarray data management and analysis [36], (iii) the mixture model method implemented in FASABI.

Storing normalized and analyzed data in our relational model allows flexible comparisons across different chips at the level of individual genes.

Quality control

Quality control reports are generated using affyQCReport- an R package that generates quality control reports for affymetrix array data [37, 38] for all experiments where .CEL files are available. In addition, we calculate a coefficient of variation of individual Probe Sets between the replicates, which provides a direct estimate of the quality between technical replicates.

Construction and content

Experiment and Sample details

The RETINOBASE home page presents the user with a list of all experiments available and also provides access to experimental details such as title, short description etc. “Sample details” option (Figure 1) gives details about sample description, organism, tissue, treatment, strain specific information and the array used for hybridisation for a given experiment.

Querying the Database

RETINOBASE has three different querying modules, “Gene Information”, “Raw Data System Analysis” and “Fold change system Analysis”.

Gene information module

The “Gene Information” module offers three different query options - “Gene Query”, “Ortholog Query” and “Blast Query”. Using these one can access information such as chromosomal location, linked retinal diseases, cellular localization, and gene ontologies for a given gene. Furthermore, gene details returned from these queries are linked to external databases such as GeneCards (http://www.genecards.org/index.shtml), NCBI (http://www.ncbi.nlm.nih.gov/), ADAPT mapping viewer [39], UniGene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene) and UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgGateway) that yield more information. (Figure 2). "Gene Query" and "Ortholog Query" accept as input the gene name, symbol, Affymterix Probe Set ID, Refseq or Unigene IDs, where as "Blast Query" accepts sequences in FASTA format. "Ortholog Query" is useful in cross-referencing probe sets between different Affymetrix Genechip arrays. The data on reference sequence similarity are taken from HomoloGene and cross-referenced. In addition, the raw data and cluster information for a given gene (cluster number, software used for clustering and information about other genes present in the same cluster) for all experiments can be obtained through the "Gene Query" (Figure 2).

Raw Data System Analysis module

This module has "Data and Cluster Query" options and "Data visualization" which is both a query and visualization option. "Data Query" (Figure 3) provides gene expression information at the level of signal intensities for single or multiple genes in one more experiments. "Cluster Query" (Figure 3) - unique to RETINOBASE, provides information of expression patterns of related genes across varied conditions and genetic backgrounds. It also identifies any two given genes in the same cluster in one or more experiments. Apart from the above mentioned query options, RETINOBASE also have a user-friendly transcriptomic data visualization tool that was developed to allow retinal biologists to graphically analyse gene expression profiles across all the experiments. A user can choose the experiment, chip, gene and analysis software to be used in a step-by-step process, following which the related samples can be labelled and organized for an easy comparison through histograms or radar-graph representations (Figure 4). This step-by-step process effectively increases querying speed, which in turn allows faster retrieval of specific data from large volumes of gene expression information. Additional information concerning the number of Probe Sets for a gene on a given chip, the normalization software used to obtain the signal intensities and the quality control report of the experiment are provided.

Fold change System Analysis module

Gene expression information at the level of fold change is provided for single or multiple genes in one or more experiments. In addition, "Ratio Query" supports a specialized query that permits retrieval of all genes from one or more experiments that has fold change greater and/or less than a given criteria.

Downloading results and User manual

In order to allow users to further compare and interpret data, results from all querying modules available in RETINOBASE can be downloaded in the comma separated value (.CSV) file format using the "Download results" option.

A user manual that provides a detailed description of the utilities in RETINOBASE is available at ***http://alnitak.u-strasbg.fr/RetinoBase/usermanual_retinobase.doc***.

Future direction

RETINOBASE is under constant development, including addition of new experiments when available. In addition, data from proprietary experiments can be accessed on approval by individual researchers and will be made generally available after publication. Several functional enhancements are also planned for the future. We will continue to refine and update RETINOBASE with respect to data retrieval, mining and visualization options. Direct upload and meta-analysis options will also be provided in the future.

Conclusions

RETINOBASE has been developed to store, analyse, visualize and compare retinal-related data in order to provide insights into retinal gene expression in various mouse models and other organisms under diverse conditions. Our database, with different types of query options and powerful visualization tools, allows comprehensive analysis of biological mechanisms/pathways of the retina in normal and under diseased conditions. The usefulness of the system should be further enhanced as new experiments are integrated.

Availability and requirements

The RETINOBASE can be accessed at ***http://alnitak.u-strasbg.fr/RetinoBase/.*** All users must register (name and email address) to obtain a username and password.

Authors' contributions

RK is involved in database design and development, data analysis, design of the user interface and prepared the manuscript; NG, GB and RR developed the web services and database back end; WR is involved in data analysis and helped to draft the manuscript; TL participated in the design of the user interface; OP was involved in drafting the manuscript.

Acknowledgements

We would like to thank Naomi Berdugo and Laëtitia Poidevin, who were the first users of the RETINOBASE, for their valuable suggestions. We thank Beena Patricia Jeevan Raj and Julie Thompson for proofreading of the manuscript. The work was supported by the European Retinal Research Training Network (RETNET) MRTN-CT-2003-504003, EVI-GENORET LSHG-CT-2005-512036, CNRS, INSERM and University of Louis Pasteur (ULP), Strasbourg, France.

.

References

1. Masland RH: **The fundamental plan of the retina**. *Nat Neurosci* 2001, **4**:877-886.
2. Pittler SJ, Baehr W: **Identification of a nonsense mutation in the rod photoreceptor cGMP phosphodiesterase beta-subunit gene of the rd mouse**. *Proc Natl Acad Sci U S A* 1991, **88**:8322-8326.
3. Akhmedov NB, Piriev NI, Chang B, Rapoport AL, Hawes NL, Nishina PM, Nusinowitz S, Heckenlively JR, Roderick TH, Kozak CA *et al*: **A deletion in a photoreceptor-specific nuclear receptor mRNA causes retinal degeneration in the rd7 mouse**. *Proc Natl Acad Sci U S A* 2000, **97**:5551-5556.
4. Pang JJ, Chang B, Hawes NL, Hurd RE, Davisson MT, Li J, Noorwez SM, Malhotra R, McDowell JH, Kaushal S *et al*: **Retinal degeneration 12 (rd12): a new, spontaneously arising mouse model for human Leber congenital amaurosis (LCA)**. *Mol Vis* 2005, **11**:152-162.
5. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles--database and tools update**. *Nucleic Acids Res* 2007, **35**:D760-765.
6. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M *et al*: **ArrayExpress--a public database of microarray experiments and gene expression profiles**. *Nucleic Acids Res* 2007, **35**:D747-750.
7. Kato K, Yamashita R, Matoba R, Monden M, Noguchi S, Takagi T, Nakai K: **Cancer gene expression database (CGED): a database for gene expression**

profiling with accompanying clinical information of human cancer tissues. *Nucleic Acids Res* 2005, **33**:D533-536.

8. Shah V, Sridhar S, Beane J, Brody JS, Spira A: **SIEGE: Smoking Induced Epithelial Gene Expression Database**. *Nucleic Acids Res* 2005, **33**:D573-579.
9. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A *et al*: **Large-scale analysis of the human and mouse transcriptomes**. *Proc Natl Acad Sci U S A* 2002, **99**:4465-4470.
10. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**:249-264.
11. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection**. *Proc Natl Acad Sci U S A* 2001, **98**:31-36.
12. Hubbell E, Liu WM, Mei R: **Robust estimators for expression analysis**. *Bioinformatics* 2002, **18**:1585-1592.
13. Hartigan JA WM: **A K-means clustering algorithm**. *Applied Statisticd* 1979, **28**:100-108.
14. Medvedovic M, Sivaganesan S: **Bayesian infinite mixture model based clustering of gene expression profiles**. *Bioinformatics* 2002, **18**:1194-1206.
15. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C: **BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data**. *Genome Biol* 2002, **3**:SOFTWARE0003.

16. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC *et al*: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data**. *Nat Genet* 2001, **29**:365-371.

17. Akimoto M, Cheng H, Zhu D, Brzezinski JA, Khanna R, Filippova E, Oh EC, Jing Y, Linares JL, Brooks M *et al*: **Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors**. *Proc Natl Acad Sci U S A* 2006, **103**:3890-3895.

18. Yoshida S, Mears AJ, Friedman JS, Carter T, He S, Oh E, Jing Y, Farjo R, Fleury G, Barlow C *et al*: **Expression profiling of the developing and mature Nrl-/- mouse retina: identification of retinal disease candidates and transcriptional regulatory targets of Nrl**. *Hum Mol Genet* 2004, **13**:1487-1503.

19. Chen J, Rattner A, Nathans J: **The rod photoreceptor-specific nuclear receptor Nr2e3 represses transcription of multiple cone-specific genes**. *J Neurosci* 2005, **25**:118-129.

20. Liu J, Huang Q, Higdon J, Liu W, Xie T, Yamashita T, Cheon K, Cheng C, Zuo J: **Distinct gene expression profiles and reduced JNK signaling in retinitis pigmentosa caused by RP1 mutations**. *Hum Mol Genet* 2005, **14**:2945-2958.

21. Cottet S, Michaut L, Boisset G, Schlecht U, Gehring W, Schorderet DF: **Biological characterization of gene response in Rpe65-/- mouse model of Leber's congenital amaurosis during progression of the disease**. *Faseb J* 2006, **20**:2036-2049.

22. Vazquez-Chona F, Song BK, Geisert EE, Jr.: **Temporal changes in gene expression after injury in the rat retina**. *Invest Ophthalmol Vis Sci* 2004, **45**:2737-2746.

23. Cheng H, Aleman TS, Cideciyan AV, Khanna R, Jacobson SG, Swaroop A: **In vivo function of the orphan nuclear receptor NR2E3 in establishing photoreceptor identity during mammalian retinal development**. *Hum Mol Genet* 2006, **15**:2588-2602.

24. Gerhardinger C, Costa MB, Coulombe MC, Toth I, Hoehn T, Grosu P: **Expression of acute-phase response proteins in retinal Muller cells in diabetes**. *Invest Ophthalmol Vis Sci* 2005, **46**:349-357.

25. Steele MR, Inman DM, Calkins DJ, Horner PJ, Vetter ML: **Microarray analysis of retinal gene expression in the DBA/2J model of glaucoma**. *Invest Ophthalmol Vis Sci* 2006, **47**:977-985.

26. Cameron DA, Gentile KL, Middleton FA, Yurco P: **Gene expression profiles of intact and regenerating zebrafish retina**. *Mol Vis* 2005, **11**:775-791.

27. Abou-Sleymane G, Chalmel F, Helmlinger D, Lardenois A, Thibault C, Weber C, Merienne K, Mandel JL, Poch O, Devys D *et al*: **Polyglutamine expansion causes neurodegeneration by altering the neuronal differentiation program**. *Hum Mol Genet* 2006, **15**:691-703.

28. Kirwan RP, Leonard MO, Murphy M, Clark AF, O'Brien CJ: **Transforming growth factor-beta-regulated gene transcription and protein expression in human GFAP-negative lamina cribrosa cells**. *Glia* 2005, **52**:309-324.

29. Zhang J, Gray J, Wu L, Leone G, Rowan S, Cepko CL, Zhu X, Craft CM, Dyer MA: **Rb regulates proliferation and rod photoreceptor development in the mouse retina**. *Nat Genet* 2004, **36**:351-360.

30. Leung YF, Ma P, Dowling JE: **Gene expression profiling of zebrafish embryonic retinal pigment epithelium in vivo**. *Invest Ophthalmol Vis Sci* 2007, **48**:881-890.

31. Carter TA, Greenhall JA, Yoshida S, Fuchs S, Helton R, Swaroop A, Lockhart DJ, Barlow C: **Mechanisms of aging in senescence-accelerated mice**. *Genome Biol* 2005, **6**:R48.

32. Michaut L, Flister S, Neeb M, White KP, Certa U, Gehring WJ: **Analysis of the eye developmental pathway in Drosophila using DNA microarrays**. *Proc Natl Acad Sci U S A* 2003, **100**:4024-4029.

33. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al*: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**:R80.

34. Wicker N, Dembele D, Raffelsberger W, Poch O: **Density of points clustering, application to transcriptomic data analysis**. *Nucleic Acids Res* 2002, **30**:3992-4000.

35. Soukas A, Cohen P, Socci ND, Friedman JM: **Leptin-specific patterns of gene expression in white adipose tissue**. *Genes Dev* 2000, **14**:963-980.

36. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J: **TM4 microarray software suite**. *Methods Enzymol* 2006, **411**:134-193.

37. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level**. *Bioinformatics* 2004, **20**:307-315.

38. Wilson CL, Miller CJ: **Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis**. *Bioinformatics* 2005, **21**:3683-3685.

39. Leong HS, Yates T, Wilson C, Miller CJ: **ADAPT: a database of affymetrix probesets and transcripts**. *Bioinformatics* 2005, **21**:2552-2553.

Figures

Figure 1 - RETINOBASE home page *(http://alnitak.u-strasbg.fr/RetinoBase/)*

Indicating general information such as experiment and sample details. Specific query options are shown as in the database.

Figure 2 - RETINOBASE Queries

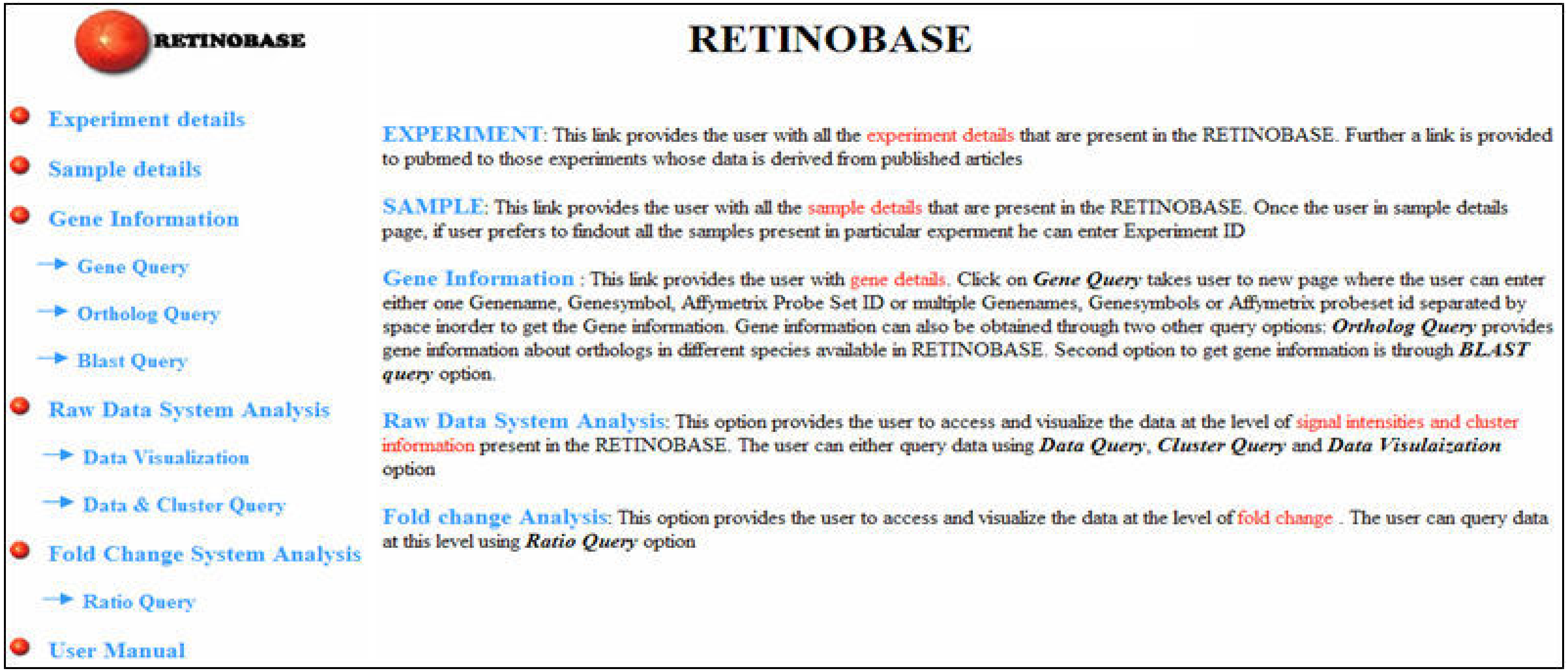
Figure legend A "Gene Query" yields information such as Unigene ID, chromosomal location, Entrez gene, expression pattern, linked diseases and gene ontology. Thick black arrow indicates that raw data and cluster information can be accessed directly from a "Gene Query" output and the dotted line indicates links to external databases.

Figure 3 - Data and Cluster Query options

Indicates Data and cluster query results for the *NRL* gene in experiment 7[17]: "Targeting GFP to new born by NRL promoter and temporal expression profiling of flow-sorted photoreceptors". From cluster data results user can obtain all genes present in the given cluster.

Figure 4 - Data visualization

Expression profile of two Probe Sets of cone-rod homeobox containing gene (CRX) in the experiment 7[17]: "Targeting GFP to new born by NRL promoter and temporal expression profiling of flow-sorted photoreceptors". Data is represented as radar plots on the top panel and as histograms in the bottom panel.



RETINOBASE

EXPERIMENT: This link provides the user with all the experiment details that are present in the RETINOBASE. Further a link is provided to pubmed to those experiments whose data is derived from published articles

SAMPLE: This link provides the user with all the sample details that are present in the RETINOBASE. Once the user in sample details page, if user prefers to findout all the samples present in particular experment he can enter Experiment ID

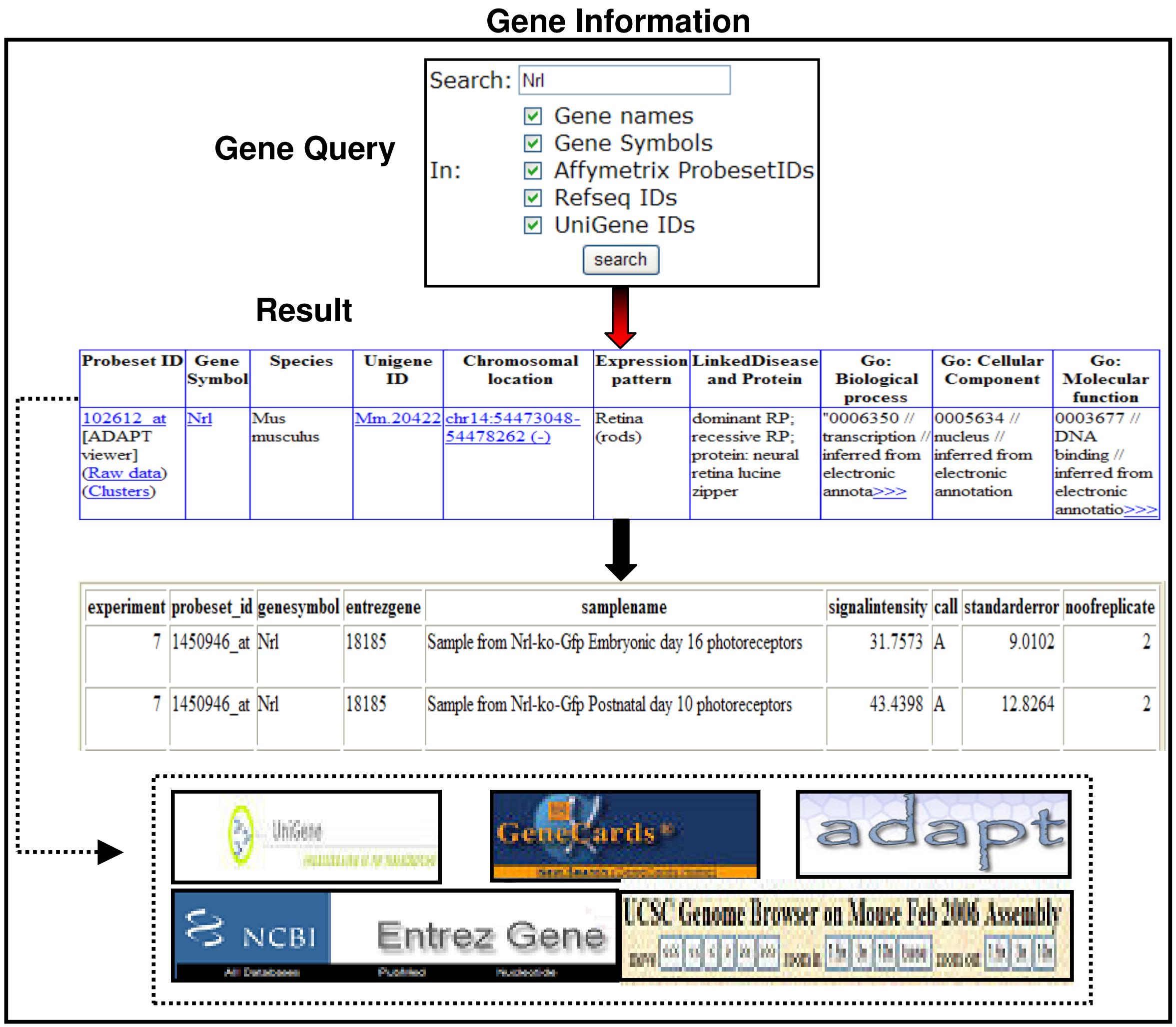
Gene Information : This link provides the user with gene details. Click on ***Gene Query*** takes user to new page where the user can enter either one Genename, Genesymbol, Affymetrix Probe Set ID or multiple Genenames, Genesymbols or Affymetrix probeset id separated by space inorder to get the Gene information. Gene information can also be obtained through two other query options: ***Ortholog Query*** provides gene information about orthologs in different species available in RETINOBASE. Second option to get gene information is through ***BLAST query*** option.

Raw Data System Analysis: This option provides the user to access and visualize the data at the level of signal intensities and cluster information present in the RETINOBASE. The user can either query data using ***Data Query***, ***Cluster Query*** and ***Data Visulaization*** option

Fold change Analysis: This option provides the user to access and visualize the data at the level of fold change . The user can query data at this level using ***Ratio Query*** option

Figure 1

Gene Information



Gene Query

Search: Nrl

In:
☑ Gene names
☑ Gene Symbols
☑ Affymetrix ProbesetIDs
☑ Refseq IDs
☑ UniGene IDs

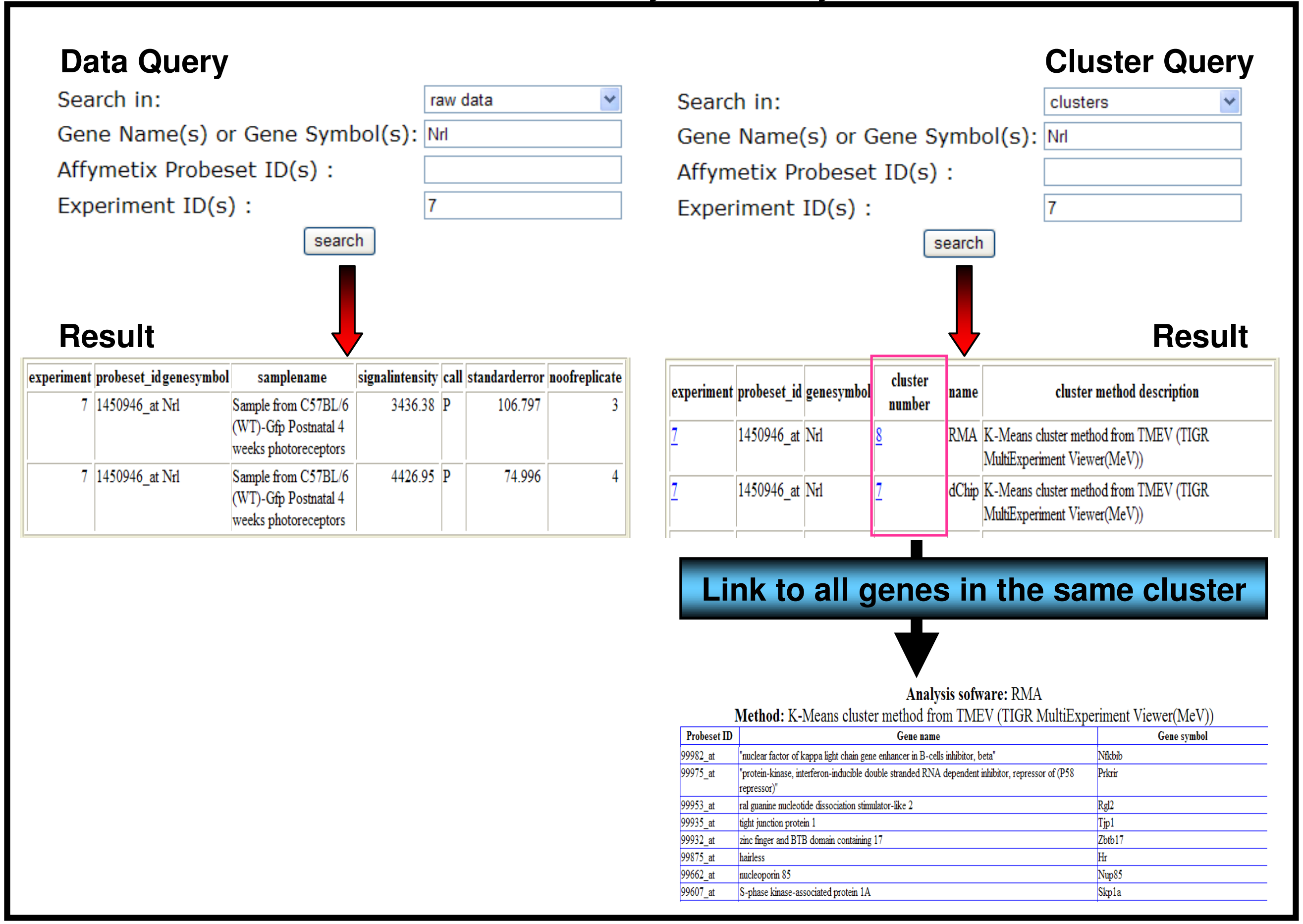
Result

Probeset ID	Gene Symbol	Species	Unigene ID	Chromosomal location	Expression pattern	LinkedDisease and Protein	Go: Biological process	Go: Cellular Component	Go: Molecular function
102612_at [ADAPT viewer] (Raw data) (Clusters)	Nrl	Mus musculus	Mm.20422	chr14:54473048-54478262 (-)	Retina (rods)	dominant RP; recessive RP; protein: neural retina lucine zipper	"0006350 // transcription // inferred from electronic annota>>>	0005634 // nucleus // inferred from electronic annotation	0003677 // DNA binding // inferred from electronic annotatio>>>

experiment	probeset_id	genesymbol	entrezgene	samplename	signalintensity	call	standarderror	noofreplicate
7	1450946_at	Nrl	18185	Sample from Nrl-ko-Gfp Embryonic day 16 photoreceptors	31.7573	A	9.0102	2
7	1450946_at	Nrl	18185	Sample from Nrl-ko-Gfp Postnatal day 10 photoreceptors	43.4398	A	12.8264	2

Figure 2

Raw Data System Analysis



Data Query

Search in: raw data

Gene Name(s) or Gene Symbol(s): Nrl

Affymetix Probeset ID(s) :

Experiment ID(s) : 7

Result

experiment	probeset_id	genesymbol	samplename	signalintensity	call	standarderror	noofreplicate
7	1450946_at	Nrl	Sample from C57BL/6 (WT)-Gfp Postnatal 4 weeks photoreceptors	3436.38	P	106.797	3
7	1450946_at	Nrl	Sample from C57BL/6 (WT)-Gfp Postnatal 4 weeks photoreceptors	4426.95	P	74.996	4

Cluster Query

Search in: clusters

Gene Name(s) or Gene Symbol(s): Nrl

Affymetix Probeset ID(s) :

Experiment ID(s) : 7

Result

experiment	probeset_id	genesymbol	cluster number	name	cluster method description
7	1450946_at	Nrl	8	RMA	K-Means cluster method from TMEV (TIGR MultiExperiment Viewer(MeV))
7	1450946_at	Nrl	7	dChip	K-Means cluster method from TMEV (TIGR MultiExperiment Viewer(MeV))

Link to all genes in the same cluster

Analysis sofware: RMA

Method: K-Means cluster method from TMEV (TIGR MultiExperiment Viewer(MeV))

Probeset ID	Gene name	Gene symbol
99982_at	"nuclear factor of kappa light chain gene enhancer in B-cells inhibitor, beta"	Nfkbib
99975_at	"protein-kinase, interferon-inducible double stranded RNA dependent inhibitor, repressor of (P58 repressor)"	Prkrir
99953_at	ral guanine nucleotide dissociation stimulator-like 2	Rgl2
99935_at	tight junction protein 1	Tjp1
99932_at	zinc finger and BTB domain containing 17	Zbtb17
99875_at	hairless	Hr
99662_at	nucleoporin 85	Nup85
99607_at	S-phase kinase-associated protein 1A	Skp1a

Figure 3

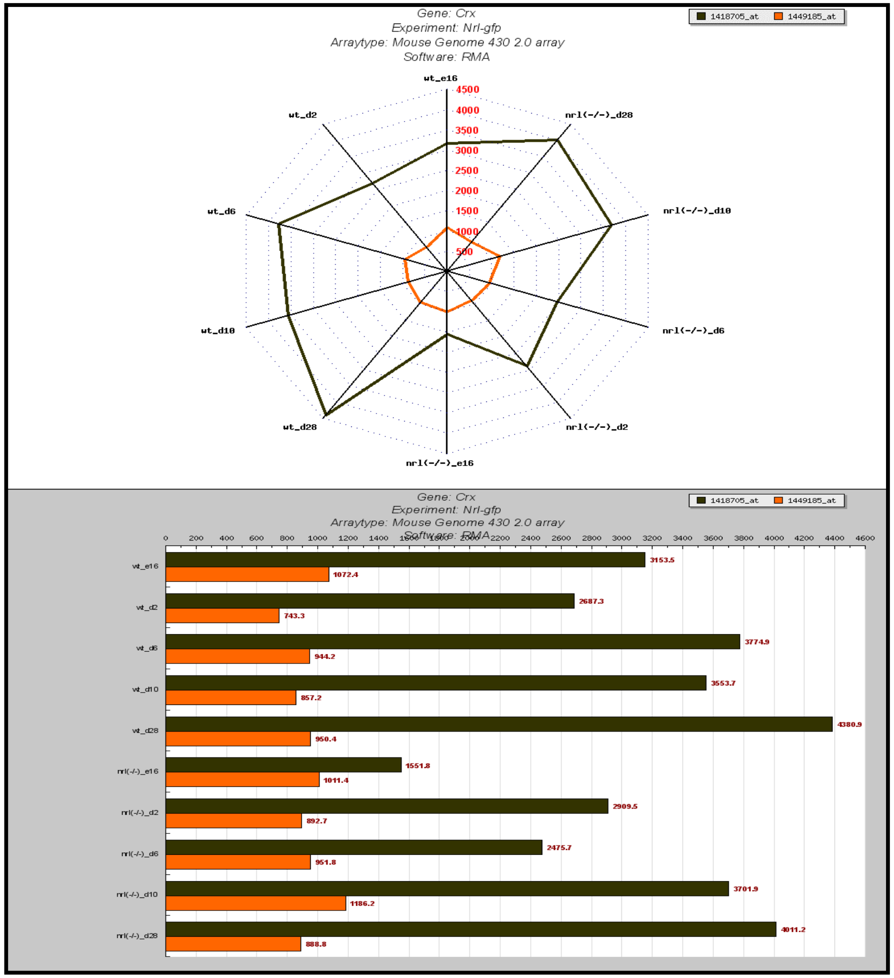

Gene: Crx
Experiment: Nrl-gfp
Arraytype: Mouse Genome 430 2.0 array
Software: RMA
■ 1418705_at ■ 1449185_at
wt_e16
nrl(-/-)_d28
nrl(-/-)_d10
nrl(-/-)_d6
nrl(-/-)_d2
nrl(-/-)_e16
wt_d28
wt_d10
wt_d6
wt_d2
4500
4000
3500
3000
2500
2000
1500
1000
500
Gene: Crx
Experiment: Nrl-gfp
Arraytype: Mouse Genome 430 2.0 array
Software: RMA
■ 1418705_at ■ 1449185_at
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
3600
3800
4000
4200
4400
4600
wt_e16
3153.5
1072.4
wt_d2
2687.3
743.3
wt_d6
3774.9
944.2
wt_d10
3553.7
857.2
wt_d28
4380.9
950.4
nrl(-/-)_e16
1551.8
1011.4
nrl(-/-)_d2
2909.5
892.7
nrl(-/-)_d6
2475.7
951.8
nrl(-/-)_d10
3701.9
1186.2
nrl(-/-)_d28
4011.2
888.8

Figure 4

Conclusions and perspectives

The post-genomic era is characterized by a flood of high volumes of data that needs to be managed efficiently. This new scientific paradigm in biology is mainly linked to two major advances, namely automatic complete genome sequencing and the emergence of high throughput technologies. This offers the opportunity to study simultaneously not only the expression of thousands of genes, proteins or metabolites but also their interactions at the cellular, tissular or organismal levels during almost all life periods. In parallel, informatic advancements have helped biologists to extract pertinent knowledge from the large volumes of data generated, through efficient interconnected analysis, storage and querying systems. Clearly, in the future, these developments will not only help to frame and guide biological hypotheses and results analysis, but also to model complex biological systems.

Unfortunately, data from these high throughput experiments are frequently associated with a high proportion of noise, which slows down their efficient exploitation and means that they need to be processed using a variety of statistical procedures.

In this context, during my thesis, I had the opportunity to participate in the development of two new algorithms to analyse high throughput data and to create a new relational database RETINOBASE to store, query and visualize processed gene expression data related to retinal tissue.

These new algorithms work at two different levels of data analysis, one at the level of clustering and the other at the level of meta-analysis. In these two projects, I was mainly involved at the level of data analysis through a biological application. In the first project concerning the Dirichlet distribution, Nicolas Wicker of our laboratory developed a new algorithm to estimate the parameters of the Dirichlet distribution using a Maximum Likelihood Approximation (MLA) method. After validation on synthetic data, we decided to test this method on biological data by clustering human proteins based on their amino acid composition. The clusters obtained by the MLA method were compared to the clusters obtained by the classical moments method in a mixture model background, revealing that the clusters unique to MLA were of added value with functional significance.

The second project concerned meta-analysis of gene expression data by a Multi-Dimensional Fitting (MDF) approach. Our laboratory in collaboration with Claude Berge and Nicolas Froloff developed this new method, which transforms datasets of one target matrix to fit distances computed on a reference matrix through a mean square objective function. After validation on synthetic data, we applied this method to gene expression data through the creation of two distinct and heterogeneous target matrices from non-neuronal tissues of lung, muscle and testis origin or from neuronal tissues of brain and retinal origin obtained from RETINOBASE, taking as reference matrix the gene expression datasets from brain. After MDF treatment, we further performed functional analysis of the target matrix genes that required the least fitting to the brain reference matrix. We observed statistically significant enrichment of biological functions that are specific to each target matrix with (i) neuron development and differentiation for the neuronal target matrix and (ii) gamete formation and spermatogenesis for the non-neuronal target matrix. These results clearly show that MDF may represent, in the future, a powerful class of methods to analyse distinct types of heterogeneous data embedded in matrices. At a more pragmatic level, we plan to apply MDF on the retinal gene expression data available in RETINOBASE to identify rapidly and systematically genes and biological processes common to experiments resulting from diverse biological questions.

The third and main project of my thesis was dedicated to the design and implementation of a new relational database RETINOBASE, conceived as a central integrative database and analysis tool for the retinal gene expression data. This major effort was developed in the framework of a European FP6 program (RETNET for "European Retinal Research Training Network") gathering most of the European retinal experts. During the initial phase of this project, one major question was related to the choice between Relational or Object-oriented DataBase Management Systems, namely RDBMS and ODBMS. Indeed, while RDBMS allows efficient and rapid handling and querying of the large volumes of datasets produced by transcriptomics data, ODBMS offers enhanced facilities for the creation of ergonomic interfaces and visualisation tools amenable to non-computer specialist users. We decided to take advantage of RDBMS 's flexibility and to develop RETINOBASE as a relational database with specific architecture and schema allowing not only efficient storage and querying but also the implementation of visualisation tools dedicated to simplified and readable display of gene expression data. This strategy has a strong impact on the overall structure of the database requiring the creation of several specialized modules or tables encompassing for example, the major steps of a transcriptomics experiment or the display on-the-fly of the gene expression level on specially designed "radar plots". Nevertheless, these specific developments do not impair the overall performance of our database which allows rapid and user-friendly access and querying to the 30 million gene expression values integrated in RETINOBASE.

A second major aspect of RETINOBASE is linked to our goal of developing a relational database as a research tool allowing the study of the advantages and inconveniences of the various transcriptomics data treatment strategies. This objective had also numerous consequences on the database architecture, notably through the creation or modification of various tables dispersed in the overall relational schema of the database. For one single transcriptomics data type, these specific adjustments allow the maintenance and querying of up to 6 values resulting from the distinct normalization and clustering treatments automatically performed in RETINOBASE. Even though, during my thesis, I did not have the opportunity to fully exploit the research tool aspects of RETINOBASE, I was able to verify that the relational schema developed and the querying system implemented allow an efficient retrieval, cross-comparison and exploitation of the various data. This was realized through the automated selection of a retinal gene set retained for the design of RetChip, a dedicated oligonucleotide microarray to study mouse retinal development and degeneration. This simple exploitation of the RETINOBASE as a research tool allowed us to validate the efficiency and rapidity of the diverse cross-comparisons performed during the selection process. Nevertheless, it is clear that, in the future, more effort will be carried on this aspect, notably through the in-depth study of the impact of specific combinations of normalization and clustering treatments on the final transcriptomics results. These future analyses will not only invoke the present data and tools available in RETINOBASE but will also take advantage of additional transcriptomics data, notably from brain origin, and of the high throughput analysis programs being developed in our laboratory.

Finally, we strongly believe that, through the efficient accumulation in an appropriate network of gene expression data resulting from a variety of treatments, experiments, tissues or animal models, RETINOBASE represents a new and effective instrument to study and elucidate the various events underlying the normal or pathological retinal and neuronal tissue development.

References

References

Acland, G.M., Aguirre, G.D., Bennett, J., Aleman, T.S., Cideciyan, A.V., Bennicelli, J., Dejneka, N.S., Pearce-Kelling, S.E., Maguire, A.M., Palczewski, K., Hauswirth, W.W. and Jacobson, S.G.: Long-term restoration of rod and cone vision by single dose rAAV-mediated gene transfer to the retina in a canine model of childhood blindness. Mol Ther 12 (2005) 1072-82.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F. and et al.: Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252 (1991) 1651-6.

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O. and et al.: Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377 (1995) 3-174.

Akaike, H.: A new look at the statistical model identification. Automatic Control, IEEE Transaction on 19 (1974) 716-713.

Akhmedov, N.B., Piriev, N.I., Chang, B., Rapoport, A.L., Hawes, N.L., Nishina, P.M., Nusinowitz, S., Heckenlively, J.R., Roderick, T.H., Kozak, C.A., Danciger, M., Davisson, M.T. and Farber, D.B.: A deletion in a photoreceptor-specific nuclear receptor mRNA causes retinal degeneration in the rd7 mouse. Proc Natl Acad Sci U S A 97 (2000) 5551-6.

Alberts, B.: Molecular biology of the cell, 4th ed. Garland Science, New York, 2002.

Alwine, J.C., Kemp, D.J. and Stark, G.R.: Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc Natl Acad Sci U S A 74 (1977) 5350-4.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. and Yeh, L.S.: UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32 (2004) D115-9.

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.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25 (2000) 25-9.

Augen, J.: Information technology to the rescue! Nat Biotechnol 19 Suppl (2001) BE39-40.

Barrett, T. and Edgar, R.: Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol 411 (2006) 352-69.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M. and Edgar, R.: NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res (2006).

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M. and Edgar, R.: NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res 35 (2007) D760-5.

Batten, M.L., Imanishi, Y., Tu, D.C., Doan, T., Zhu, L., Pang, J., Glushakova, L., Moise, A.R., Baehr, W., Van Gelder, R.N., Hauswirth, W.W., Rieke, F. and Palczewski, K.: Pharmacological and rAAV gene therapy rescue of visual functions in a blind mouse model of Leber congenital amaurosis. PLoS Med 2 (2005) e333.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L.: GenBank. Nucleic Acids Res 35 (2007) D21-5.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res 28 (2000) 235-42.

Blackshaw, S., Fraioli, R.E., Furukawa, T. and Cepko, C.L.: Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. Cell 107 (2001) 579-89.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31 (2003) 365-70.

Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19 (2003) 185-93.

Bortoluzzi, S., d'Alessi, F. and Danieli, G.A.: A novel resource for the study of genes expressed in the adult human retina. Invest Ophthalmol Vis Sci 41 (2000) 3305-8.

Bowes Rickman, C., Ebright, J.N., Zavodni, Z.J., Yu, L., Wang, T., Daiger, S.P., Wistow, G., Boon, K. and Hauser, M.A.: Defining the human macula transcriptome and candidate retinal disease genes using EyeSAGE. Invest Ophthalmol Vis Sci 47 (2006) 2305-16.

Broberg, P.: Statistical methods for ranking differentially expressed genes. Genome Biol 4 (2003) R41.

Cabral, A., Voskamp, P., Cleton-Jansen, A.M., South, A., Nizetic, D. and Backendorf, C.: Structural organization and regulation of the small proline-rich family of cornified envelope precursors suggest a role in adaptive barrier function. J Biol Chem 276 (2001) 19231-7.

Cedano, J., Aloy, P., Perez-Pons, J.A. and Querol, E.: Relation between amino acid composition and cellular location of proteins. J Mol Biol 266 (1997) 594-600.

Cheng, G. and Porter, J.D.: Transcriptional profile of rat extraocular muscle by serial analysis of gene expression. Invest Ophthalmol Vis Sci 43 (2002) 1048-58.

Cheung, K.H., Yip, K.Y., Smith, A., Deknikker, R., Masiar, A. and Gerstein, M.: YeastHub: a semantic web use case for integrating data in the life sciences domain. Bioinformatics 21 Suppl 1 (2005) i85-96.

Cho, R.J., Huang, M., Campbell, M.J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S.J., Davis, R.W. and Lockhart, D.J.: Transcriptional regulation and function during the human cell cycle. Nat Genet 27 (2001) 48-54.

Chowers, I., Gunatilaka, T.L., Farkas, R.H., Qian, J., Hackam, A.S., Duh, E., Kageyama, M., Wang, C., Vora, A., Campochiaro, P.A. and Zack, D.J.: Identification of novel genes preferentially expressed in the retina using a custom human retina cDNA microarray. Invest Ophthalmol Vis Sci 44 (2003a) 3732-41.

Chowers, I., Liu, D., Farkas, R.H., Gunatilaka, T.L., Hackam, A.S., Bernstein, S.L., Campochiaro, P.A., Parmigiani, G. and Zack, D.J.: Gene expression variation in the adult human retina. Hum Mol Genet 12 (2003b) 2881-93.

Crick, F.H.: On protein synthesis. Symp Soc Exp Biol 12 (1958) 138-63.

Dalke, C. and Graw, J.: Mouse mutants as models for congenital retinal disorders. Exp Eye Res 81 (2005) 503-12.

Deloukas, P., Schuler, G.D., Gyapay, G., Beasley, E.M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T.C., McKusick, K.B., Beckmann, J.S., Bentolila, S., Bihoreau, M., Birren, B.B., Browne, J., Butler, A., Castle, A.B., Chiannilkulchai, N., Clee, C., Day, P.J., Dehejia, A., Dibling, T., Drouot, N., Duprat, S., Fizames, C., Fox, S., Gelling, S., Green, L., Harrison, P., Hocking, R., Holloway, E., Hunt, S., Keil, S., Lijnzaad, P., Louis-Dit-Sully, C., Ma, J., Mendis, A., Miller, J., Morissette, J., Muselet, D., Nusbaum, H.C., Peck, A., Rozen, S., Simon, D., Slonim, D.K., Staples, R., Stein, L.D., Stewart, E.A., Suchard, M.A., Thangarajah, T., Vega-Czarny, N., Webber, C., Wu, X., Hudson, J., Auffray, C., Nomura, N., Sikela, J.M., Polymeropoulos, M.H., James, M.R., Lander, E.S., Hudson, T.J., Myers, R.M., Cox, D.R., Weissenbach, J., Boguski, M.S. and Bentley, D.R.: A physical map of 30,000 human genes. Science 282 (1998) 744-6.

Delsuc, F., Brinkmann, H. and Philippe, H.: Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6 (2005) 361-75.

Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A.: DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4 (2003) P3.

Dharmaraj, S., Li, Y., Robitaille, J.M., Silva, E., Zhu, D., Mitchell, T.N., Maltby, L.P., Baffoe-Bonnie, A.B. and Maumenee, I.H.: A novel locus for Leber congenital amaurosis maps to chromosome 6q. Am J Hum Genet 66 (2000) 319-26.

Diaz, E., Yang, Y.H., Ferreira, T., Loh, K.C., Okazaki, Y., Hayashizaki, Y., Tessier-Lavigne, M., Speed, T.P. and Ngai, J.: Analysis of gene expression in the developing mouse retina. Proc Natl Acad Sci U S A 100 (2003) 5491-6.

Draghici, S.: Statistical intelligence: effective analysis of high-density microarray data. Drug Discov Today 7 (2002) S55-63.

Dredge, B.K., Polydorides, A.D. and Darnell, R.B.: The splice of life: alternative splicing and neurological disease. Nat Rev Neurosci 2 (2001) 43-50.

Efron B, T.R., Storey JD: Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 96 (2001) 1151-60.

Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M.: The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 6 (2005) R44.

Eisen, M.B. and Brown, P.O.: DNA arrays for analysis of gene expression. Methods Enzymol 303 (1999) 179-205.

Elnitski, L., Jin, V.X., Farnham, P.J. and Jones, S.J.: Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res 16 (2006) 1455-64.

Flanders, M., Lapointe, M.L., Brownstein, S. and Little, J.M.: Keratoconus and Leber's congenital amaurosis: a clinicopathological correlation. Can J Ophthalmol 19 (1984) 310-4.

Freund, C.L., Wang, Q.L., Chen, S., Muskat, B.L., Wiles, C.D., Sheffield, V.C., Jacobson, S.G., McInnes, R.R., Zack, D.J. and Stone, E.M.: De novo mutations in the CRX homeobox gene associated with Leber congenital amaurosis. Nat Genet 18 (1998) 311-2.

Friend, S.H., Bernards, R., Rogelj, S., Weinberg, R.A., Rapaport, J.M., Albert, D.M. and Dryja, T.P.: A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. Nature 323 (1986) 643-6.

Gao, J., Cheon, K., Nusinowitz, S., Liu, Q., Bei, D., Atkins, K., Azimi, A., Daiger, S.P., Farber, D.B., Heckenlively, J.R., Pierce, E.A., Sullivan, L.S. and Zuo, J.: Progressive photoreceptor degeneration, outer segment dysplasia, and rhodopsin mislocalization in mice with targeted disruption of the retinitis pigmentosa-1 (Rp1) gene. Proc Natl Acad Sci U S A 99 (2002) 5698-703.

Garneau, N.L., Wilusz, J. and Wilusz, C.J.: The highways and byways of mRNA decay. Nat Rev Mol Cell Biol 8 (2007) 113-26.

Garnis, C., Buys, T.P. and Lam, W.L.: Genetic alteration and gene expression modulation during cancer progression. Mol Cancer 3 (2004) 9.

Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A.: affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20 (2004) 307-15.

Gayen, A.K.: The distribution of student's t in random samples of any size drawn from non-normal universes. Biometrika 36 (1949) 353-69.

Ge, H., Walhout, A.J. and Vidal, M.: Integrating 'omic' information: a bridge between genomics and systems biology. Trends Genet 19 (2003) 551-60.

Geer, R.C. and Sayers, E.W.: Entrez: making use of its power. Brief Bioinform 4 (2003) 179-84.

Gehrig, A., Langmann, T., Horling, F., Janssen, A., Bonin, M., Walter, M., Poths, S. and Weber, B.H.: Genome-wide expression profiling of the retinoschisin-deficient retina in early postnatal mouse development. Invest Ophthalmol Vis Sci 48 (2007) 891-900.

Gehrs, K.M., Anderson, D.H., Johnson, L.V. and Hageman, G.S.: Age-related macular degeneration--emerging pathogenetic and therapeutic concepts. Ann Med 38 (2006) 450-71.

Gentle, A., Anastasopoulos, F. and McBrien, N.A.: High-resolution semi-quantitative real-time PCR without the use of a standard curve. Biotechniques 31 (2001) 502, 504-6, 508.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y. and Zhang, J.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5 (2004) R80.

Gerber, S., Perrault, I., Hanein, S., Barbet, F., Ducroq, D., Ghazi, I., Martin-Coignard, D., Leowski, C., Homfray, T., Dufier, J.L., Munnich, A., Kaplan, J. and Rozet, J.M.: Complete exon-intron structure of the RPGR-interacting protein (RPGRIP1) gene allows the identification of mutations underlying Leber congenital amaurosis. Eur J Hum Genet 9 (2001) 561-71.

Gieser, L. and Swaroop, A.: Expressed sequence tags and chromosomal localization of cDNA clones from a subtracted retinal pigment epithelium library. Genomics 13 (1992) 873-6.

Glass, G.V.: The meta-analysis of research. Review of Research in Education (1977) 351-379.

Goldsmith, P. and Harris, W.A.: The zebrafish as a tool for understanding the biology of visual disorders. Semin Cell Dev Biol 14 (2003) 11-8.

Gottsch, J.D., Bowers, A.L., Margulies, E.H., Seitzman, G.D., Kim, S.W., Saha, S., Jun, A.S., Stark, W.J. and Liu, S.H.: Serial analysis of gene expression in the corneal endothelium of Fuchs' dystrophy. Invest Ophthalmol Vis Sci 44 (2003) 594-9.

Graveley, B.R.: Alternative splicing: increasing diversity in the proteomic world. Trends Genet 17 (2001) 100-7.

Gu, S.M., Thompson, D.A., Srikumari, C.R., Lorenz, B., Finckh, U., Nicoletti, A., Murthy, K.R., Rathmann, M., Kumaramanickavel, G., Denton, M.J. and Gal, A.: Mutations in RPE65 cause autosomal recessive childhood-onset severe retinal dystrophy. Nat Genet 17 (1997) 194-7.

Hannon, G.J. and Rossi, J.J.: Unlocking the potential of the human genome with RNA interference. Nature 431 (2004) 371-8.

Hartong, D.T., Berson, E.L. and Dryja, T.P.: Retinitis pigmentosa. Lancet 368 (2006) 1795-809.

Hayes, P.D., Schmitt, K., Jones, H.B., Gyapay, G., Weissenbach, J. and Goodfellow, P.N.: Regional assignment of human ESTs by whole-genome radiation hybrid mapping. Mamm Genome 7 (1996) 446-50.

Heegaard, S., Rosenberg, T., Preising, M., Prause, J.U. and Bek, T.: An unusual retinal vascular morphology in connection with a novel AIPL1 mutation in Leber's congenital amaurosis. Br J Ophthalmol 87 (2003) 980-3.

Hendrickson, A.E. and Yuodelis, C.: The morphological development of the human fovea. Ophthalmology 91 (1984) 603-12.

Hentze, M.W. and Kulozik, A.E.: A perfect message: RNA surveillance and nonsense-mediated decay. Cell 96 (1999) 307-10.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S.G., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C. and Apweiler, R.: The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol 22 (2004) 177-83.

Hernandez T, K.S.: Intergration of biological sources:current systems and challenges ahead. ACM SIGMOD 33 (2004) 51-60.

Hubbell, E., Liu, W.M. and Mei, R.: Robust estimators for expression analysis. Bioinformatics 18 (2002) 1585-92.

Humphries, M.M., Rancourt, D., Farrar, G.J., Kenna, P., Hazel, M., Bush, R.A., Sieving, P.A., Sheils, D.M., McNally, N., Creighton, P., Erven, A., Boros, A., Gulya, K., Capecchi, M.R. and Humphries, P.: Retinopathy induced in mice by targeted disruption of the rhodopsin gene. Nat Genet 15 (1997) 216-9.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P.: Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31 (2003a) e15.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4 (2003b) 249-64.

Jansen, R.P.: mRNA localization: message on the move. Nat Rev Mol Cell Biol 2 (2001) 247-56.

Jensen, L.J., Gupta, R., Staerfeldt, H.H. and Brunak, S.: Prediction of human protein function according to Gene Ontology categories. Bioinformatics 19 (2003) 635-42.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M.: From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34 (2006) D354-7.

Kato, K., Yamashita, R., Matoba, R., Monden, M., Noguchi, S., Takagi, T. and Nakai, K.: Cancer gene expression database (CGED): a database for gene expression profiling with accompanying clinical information of human cancer tissues. Nucleic Acids Res 33 (2005) D533-6.

Keeler, C.E.: The Inheritance of a Retinal Abnormality in White Mice. Proc Natl Acad Sci U S A 10 (1924) 329-33.

Kennan, A., Aherne, A., Palfi, A., Humphries, M., McKee, A., Stitt, A., Simpson, D.A., Demtroder, K., Orntoft, T., Ayuso, C., Kenna, P.F., Farrar, G.J. and Humphries, P.: Identification of an IMPDH1 mutation in autosomal dominant retinitis pigmentosa (RP10) revealed following comparative microarray analysis of transcripts derived from retinas of wild-type and Rho(-/-) mice. Hum Mol Genet 11 (2002) 547-57.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D.: The human genome browser at UCSC. Genome Res 12 (2002) 996-1006.

Kerr, M.K., Martin, M. and Churchill, G.A.: Analysis of variance for gene expression microarray data. J Comput Biol 7 (2000) 819-37.

Knudson, A.G., Jr.: Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A 68 (1971) 820-3.

Kornberg, R.D.: Eukaryotic transcriptional control. Trends Cell Biol 9 (1999) M46-9.

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P.E. and Berman, H.M.: The RCSB PDB information portal for structural genomics. Nucleic Acids Res 34 (2006) D302-5.

Kroll, A.J. and Kuwabara, T.: Electron Microscopy of a Retinal Abiotrophy. Arch Ophthalmol 71 (1964) 683-90.

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz, C., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Pastor, M.P., Plaister, S., Sobhany, S., Stoehr, P., Vaughan, R., Wu, D., Zhu, W. and Apweiler, R.: EMBL Nucleotide Sequence Database in 2006. Nucleic Acids Res 35 (2007) D16-20.

Le Brigand, K. and Barbry, P.: Mediante: a web-based microarray data manager. Bioinformatics 23 (2007) 1304-6.

Lem, J., Krasnoperova, N.V., Calvert, P.D., Kosaras, B., Cameron, D.A., Nicolo, M., Makino, C.L. and Sidman, R.L.: Morphological, physiological, and biochemical changes in rhodopsin knockout mice. Proc Natl Acad Sci U S A 96 (1999) 736-41.

Leong, H.S., Yates, T., Wilson, C. and Miller, C.J.: ADAPT: a database of affymetrix probesets and transcripts. Bioinformatics 21 (2005) 2552-3.

Leontis, N.B., Altman, R.B., Berman, H.M., Brenner, S.E., Brown, J.W., Engelke, D.R., Harvey, S.C., Holbrook, S.R., Jossinet, F., Lewis, S.E., Major, F., Mathews, D.H., Richardson, J.S., Williamson, J.R. and Westhof, E.: The RNA Ontology Consortium: an open invitation to the RNA community. Rna 12 (2006) 533-41.

Li, C. and Wong, W.H.: Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 98 (2001) 31-6.

Lin, D.M., Yang, Y.H., Scolnick, J.A., Brunet, L.J., Marsh, H., Peng, V., Okazaki, Y., Hayashizaki, Y., Speed, T.P. and Ngai, J.: Spatial patterns of gene expression in the olfactory bulb. Proc Natl Acad Sci U S A 101 (2004) 12718-23.

Lorenz, B., Gyurus, P., Preising, M., Bremser, D., Gu, S., Andrassi, M., Gerth, C. and Gal, A.: Early-onset severe rod-cone dystrophy in young children with RPE65 mutations. Invest Ophthalmol Vis Sci 41 (2000) 2735-42.

Lotery, A.J., Jacobson, S.G., Fishman, G.A., Weleber, R.G., Fulton, A.B., Namperumalsamy, P., Heon, E., Levin, A.V., Grover, S., Rosenow, J.R., Kopp, K.K., Sheffield, V.C. and Stone, E.M.: Mutations in the CRB1 gene cause Leber congenital amaurosis. Arch Ophthalmol 119 (2001) 415-20.

Lyubarsky, A.L., Falsini, B., Pennesi, M.E., Valentini, P. and Pugh, E.N., Jr.: UV- and midwave-sensitive cone-driven retinal responses of the mouse: a possible phenotype for coexpression of cone photopigments. J Neurosci 19 (1999) 442-55.

Mackiewicz, M., Shockley, K.R., Romer, M.A., Galante, R.J., Zimmerman, J.E., Naidoo, N., Baldwin, D.A., Jensen, S.T., Churchill, G.A. and Pack, A.I.: Macromolecule biosynthesis: a key function of sleep. Physiol Genomics 31 (2007) 441-57.

Madden, S.L., Galella, E.A., Zhu, J., Bertelsen, A.H. and Beaudry, G.A.: SAGE transcript profiles for p53-dependent growth regulation. Oncogene 15 (1997) 1079-85.

Malone, K., Sohocki, M.M., Sullivan, L.S. and Daiger, S.P.: Identifying and mapping novel retinal-expressed ESTs from humans. Mol Vis 5 (1999) 5.

Maxam, A.M. and Gilbert, W.: A new method for sequencing DNA. Proc Natl Acad Sci U S A 74 (1977) 560-4.

McCarthy, L.C., Terrett, J., Davis, M.E., Knights, C.J., Smith, A.L., Critcher, R., Schmitt, K., Hudson, J., Spurr, N.K. and Goodfellow, P.N.: A first-generation whole genome-radiation hybrid map spanning the mouse genome. Genome Res 7 (1997) 1153-61.

McLachlan, G.J. and Basford, K.E.: Mixture models : inference and applications to clustering. M. Dekker, New York, N.Y., 1988.

Meister, G. and Tuschl, T.: Mechanisms of gene silencing by double-stranded RNA. Nature 431 (2004) 343-9.

Mu, X., Zhao, S., Pershad, R., Hsieh, T.F., Scarpa, A., Wang, S.W., White, R.A., Beremand, P.D., Thomas, T.L., Gan, L. and Klein, W.H.: Gene expression in the developing mouse retina by EST sequencing and microarray analysis. Nucleic Acids Res 29 (2001) 4983-93.

Mulligan, M.K., Ponomarev, I., Hitzemann, R.J., Belknap, J.K., Tabakoff, B., Harris, R.A., Crabbe, J.C., Blednov, Y.A., Grahame, N.J., Phillips, T.J., Finn, D.A., Hoffman, P.L., Iyer, V.R., Koob, G.F. and Bergeson, S.E.: Toward understanding the genetics of alcohol drinking through transcriptome meta-analysis. Proc Natl Acad Sci U S A 103 (2006) 6368-73.

Nagarajan R, A.M., Phatak A: Database Challenges in the Intergration of Biomedical Data Sets, Proceedings of VLDB conference, Toronto, Canada, 2004.

Novina, C.D. and Sharp, P.A.: The RNAi revolution. Nature 430 (2004) 161-4.

Palmer, S., Wiegand, A.P., Maldarelli, F., Bazmi, H., Mican, J.M., Polis, M., Dewar, R.L., Planta, A., Liu, S., Metcalf, J.A., Mellors, J.W. and Coffin, J.M.: New real-time reverse transcriptase-initiated PCR assay with single-copy sensitivity for human immunodeficiency virus type 1 RNA in plasma. J Clin Microbiol 41 (2003) 4531-6.

Pang, I.H. and Clark, A.F.: Rodent models for glaucoma retinopathy and optic neuropathy. J Glaucoma 16 (2007) 483-505.

Pang, J.J., Chang, B., Hawes, N.L., Hurd, R.E., Davisson, M.T., Li, J., Noorwez, S.M., Malhotra, R., McDowell, J.H., Kaushal, S., Hauswirth, W.W., Nusinowitz, S., Thompson, D.A. and Heckenlively, J.R.: Retinal degeneration 12 (rd12): a new, spontaneously arising mouse model for human Leber congenital amaurosis (LCA). Mol Vis 11 (2005) 152-62.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U. and Brazma, A.: ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 35 (2007) D747-50.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G.G., Holloway, E., Kapushesky, M., Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-Serra, P., Sharma, A., Sansone, S. and Brazma, A.: ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 33 (2005) D553-5.

Pavlidis, P.: Using ANOVA for gene selection from microarray studies of the nervous system. Methods 31 (2003) 282-9.

Pavlidis, P., Li, Q. and Noble, W.S.: The effect of replication on gene expression microarray experiments. Bioinformatics 19 (2003) 1620-7.

Pennisi, E.: How will big pictures emerge from a sea of biological data? Science 309 (2005) 94.

Perrault, I., Rozet, J.M., Calvas, P., Gerber, S., Camuzat, A., Dollfus, H., Chatelin, S., Souied, E., Ghazi, I., Leowski, C., Bonnemaison, M., Le Paslier, D., Frezal, J., Dufier, J.L., Pittler, S., Munnich, A. and Kaplan, J.: Retinal-specific guanylate cyclase gene mutations in Leber's congenital amaurosis. Nat Genet 14 (1996) 461-4.

Petters, R.M., Alexander, C.A., Wells, K.D., Collins, E.B., Sommer, J.R., Blanton, M.R., Rojas, G., Hao, Y., Flowers, W.L., Banin, E., Cideciyan, A.V., Jacobson, S.G. and Wong, F.: Genetically engineered large animal model for studying cone photoreceptor survival and degeneration in retinitis pigmentosa. Nat Biotechnol 15 (1997) 965-70.

Philippi, S. and Kohler, J.: Addressing the problems with life-science databases for traditional uses and systems biology. Nat Rev Genet 7 (2006) 482-8.

Pittler, S.J. and Baehr, W.: Identification of a nonsense mutation in the rod photoreceptor cGMP phosphodiesterase beta-subunit gene of the rd mouse. Proc Natl Acad Sci U S A 88 (1991) 8322-6.

Quackenbush, J.: Computational analysis of microarray data. Nat Rev Genet 2 (2001) 418-27.

Raffelsberger, W., Krause, Y., Moulinier, L., Kieffer, D., Morand, A., Brino, L. and Poch, O.: RReportGenerator: Automatic reports from routine statistical analysis using R. Bioinformatics (2007).

Redmond, T.M., Yu, S., Lee, E., Bok, D., Hamasaki, D., Chen, N., Goletz, P., Ma, J.X., Crouch, R.K. and Pfeifer, K.: Rpe65 is necessary for production of 11-cis-vitamin A in the retinal visual cycle. Nat Genet 20 (1998) 344-51.

Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D. and Chinnaiyan, A.M.: Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res 62 (2002) 4427-33.

Rizzolo, L.J., Chen, X., Weitzman, M., Sun, R. and Zhang, H.: Analysis of the RPE transcriptome reveals dynamic changes during the development of the outer blood-retinal barrier. Mol Vis 13 (2007) 1259-73.

Saal, L.H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A. and Peterson, C.: BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol 3 (2002) SOFTWARE0003.

Saeed, A.I., Bhagabati, N.K., Braisted, J.C., Liang, W., Sharov, V., Howe, E.A., Li, J., Thiagarajan, M., White, J.A. and Quackenbush, J.: TM4 microarray software suite. Methods Enzymol 411 (2006) 134-93.

Samardzija, M., von Lintig, J., Tanimoto, N., Oberhauser, V., Thiersch, M., Reme, C.E., Seeliger, M., Grimm, C. and Wenzel, A.: R91W mutation in Rpe65 leads to milder early-onset retinal dystrophy due to the generation of low levels of 11-cis-retinal. Hum Mol Genet (2007).

Sanger, F., Nicklen, S. and Coulson, A.R.: DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74 (1977) 5463-7.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270 (1995) 467-70.

Schulz, H.L., Goetz, T., Kaschkoetoe, J. and Weber, B.H.: The Retinome - defining a reference transcriptome of the adult mammalian retina/retinal pigment epithelium. BMC Genomics 5 (2004) 50.

Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6 (1978) 461-464.

Semple-Rowland, S.L., Larkin, P., Bronson, J.D., Nykamp, K., Streit, W.J. and Baehr, W.: Characterization of the chicken GCAP gene array and analyses of GCAP1, GCAP2, and GC1 gene expression in normal and rd chicken pineal. Mol Vis 5 (1999) 14.

Seo, J., Gordish-Dressman, H. and Hoffman, E.P.: An interactive power analysis tool for microarray hypothesis testing and generation. Bioinformatics 22 (2006) 808-14.

Shah, V., Sridhar, S., Beane, J., Brody, J.S. and Spira, A.: SIEGE: Smoking Induced Epithelial Gene Expression Database. Nucleic Acids Res 33 (2005) D573-9.

Sharon, D., Blackshaw, S., Cepko, C.L. and Dryja, T.P.: Profile of the genes expressed in the human peripheral retina, macula, and retinal pigment epithelium determined through serial analysis of gene expression (SAGE). Proc Natl Acad Sci U S A 99 (2002) 315-20.

Sharp, P.A.: RNAi and double-strand RNA. Genes Dev 13 (1999) 139-41.

Shatkin, A.J. and Manley, J.L.: The ends of the affair: capping and polyadenylation. Nat Struct Biol 7 (2000) 838-42.

Shibagaki, Y., Itoh, N., Yamada, H., Nagata, S. and Mizumoto, K.: mRNA capping enzyme. Isolation and characterization of the gene encoding mRNA guanylytransferase subunit from Saccharomyces cerevisiae. J Biol Chem 267 (1992) 9521-8.

Sinha, S., Sharma, A., Agarwal, N., Swaroop, A. and Yang-Feng, T.L.: Expression profile and chromosomal location of cDNA clones, identified from an enriched adult retina library. Invest Ophthalmol Vis Sci 41 (2000) 24-8.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S. and Haussler, D.: Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci 12 (1996) 327-45.

Smyth, G.K.: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3 (2004) Article3.

Smyth, G.K. and Speed, T.: Normalization of cDNA microarray data. Methods 31 (2003) 265-73.

Sohocki, M.M., Bowne, S.J., Sullivan, L.S., Blackshaw, S., Cepko, C.L., Payne, A.M., Bhattacharya, S.S., Khaliq, S., Qasim Mehdi, S., Birch, D.G., Harrison, W.R., Elder, F.F., Heckenlively, J.R. and Daiger, S.P.: Mutations in a new photoreceptor-pineal gene on 17p cause Leber congenital amaurosis. Nat Genet 24 (2000) 79-83.

Soukas, A., Cohen, P., Socci, N.D. and Friedman, J.M.: Leptin-specific patterns of gene expression in white adipose tissue. Genes Dev 14 (2000) 963-80.

Spellman, P.T. and Rubin, G.M.: Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol 1 (2002) 5.

Steinhoff, C. and Vingron, M.: Normalization and quantification of differential expression in gene expression microarrays. Brief Bioinform 7 (2006) 166-77.

Stockton, D.W., Lewis, R.A., Abboud, E.B., Al-Rajhi, A., Jabak, M., Anderson, K.L. and Lupski, J.R.: A novel locus for Leber congenital amaurosis on chromosome 14q24. Hum Genet 103 (1998) 328-33.

Stohr, H., Mah, N., Schulz, H.L., Gehrig, A., Frohlich, S. and Weber, B.H.: EST mining of the UniGene dataset to identify retina-specific genes. Cytogenet Cell Genet 91 (2000) 267-77.

Struhl, K.: Histone acetylation and transcriptional regulatory mechanisms. Genes Dev 12 (1998) 599-606.

Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., Patapoutian, A., Hampton, G.M., Schultz, P.G. and Hogenesch, J.B.: Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A 99 (2002) 4465-70.

Sugawara, H., Abe, T., Gojobori, T. and Tateno, Y.: DDBJ working on evaluation and classification of bacterial genes in INSDC. Nucleic Acids Res 35 (2007) D13-5.

Sullivan, T.J., Heathcote, J.G., Brazel, S.M. and Musarella, M.A.: The ocular pathology in Leber's congenital amaurosis. Aust N Z J Ophthalmol 22 (1994) 25-31.

Swain, P.K., Hicks, D., Mears, A.J., Apel, I.J., Smith, J.E., John, S.K., Hendrickson, A., Milam, A.H. and Swaroop, A.: Multiple phosphorylated isoforms of NRL are expressed in rod photoreceptors. J Biol Chem 276 (2001) 36824-30.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M.: Systematic determination of genetic network architecture. Nat Genet 22 (1999) 281-5.

Thomas, J.G., Olson, J.M., Tapscott, S.J. and Zhao, L.P.: An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res 11 (2001) 1227-36.

Thompson, D.A., Li, Y., McHenry, C.L., Carlson, T.J., Ding, X., Sieving, P.A., Apfelstedt-Sylla, E. and Gal, A.: Mutations in the gene encoding lecithin retinol acyltransferase are associated with early-onset severe retinal dystrophy. Nat Genet 28 (2001) 123-4.

Travis, G.H., Brennan, M.B., Danielson, P.E., Kozak, C.A. and Sutcliffe, J.G.: Identification of a photoreceptor-specific mRNA encoded by the gene responsible for retinal degeneration slow (rds). Nature 338 (1989) 70-3.

Tusher, V.G., Tibshirani, R. and Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98 (2001) 5116-21.

Valcarcel, J. and Green, M.R.: The SR protein family: pleiotropic functions in pre-mRNA splicing. Trends Biochem Sci 21 (1996) 296-301.

Valoczi, A., Hornyik, C., Varga, N., Burgyan, J., Kauppinen, S. and Havelda, Z.: Sensitive and specific detection of microRNAs by northern blot analysis using LNA-modified oligonucleotide probes. Nucleic Acids Res 32 (2004) e175.

van de Rijn, M. and Gilks, C.B.: Applications of microarrays to histopathology. Histopathology 44 (2004) 97-108.

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W.: Serial analysis of gene expression. Science 270 (1995) 484-7.

Waardenburg, P.J. and Schappert-Kimmijser, J.: On Various Recessive Biotypes of Leber's Congenital Amaurosis. Acta Ophthalmol (Copenh) 41 (1963) 317-20.

Watson, J.D. and Crick, F.H.: Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171 (1953) 737-8.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Miller, V., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L. and Yaschenko, E.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35 (2007) D5-12.

Wicker, N., Dembele, D., Raffelsberger, W. and Poch, O.: Density of points clustering, application to transcriptomic data analysis. Nucleic Acids Res 30 (2002) 3992-4000.

Wilson, C.L. and Miller, C.J.: Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. Bioinformatics 21 (2005) 3683-5.

Wong, M.L. and Medrano, J.F.: Real-time PCR for mRNA quantitation. Biotechniques 39 (2005) 75-85.

Wu, Z. and Irizarry, R.A.: Stochastic models inspired by hybridization theory for short oligonucleotide arrays. J Comput Biol 12 (2005) 882-93.

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. and Ruzzo, W.L.: Model-based clustering and data transformations for gene expression data. Bioinformatics 17 (2001) 977-87.

Yoshida, S., Yashar, B.M., Hiriyanna, S. and Swaroop, A.: Microarray analysis of gene expression in the aging human retina. Invest Ophthalmol Vis Sci 43 (2002) 2554-60.

Zhang, S.S., Xu, X., Liu, M.G., Zhao, H., Soares, M.B., Barnstable, C.J. and Fu, X.Y.: A biphasic pattern of gene expression during mouse retina development. BMC Dev Biol 6 (2006) 48.

Annexe

Annexe 1: Standardization procedure to convert gene expression data into proportions.

For a gene with expression levels $x_1, \ldots\ldots x_p$ where, p = number of conditions or treatment for the gene x $x_i^{'} = \frac{x_i}{\sum_{j=1}^{p} x_j}$, where $x_i^{'}$ = new coordinates of the considered gene x Dirichlet's distribution works where $\sum x_i^{'} = 1$ and $x_i^{'} > 0$ conditions only.

Annexe 2: Details about data sets used for MDF analysis.

GEO accession indicates Gene Expression Omnibus accession number; Tissue type indicates the tissue used in performing that particular microarray experiment; Number of samples indicates number of datasets present under each of the target matrices and reference matrix.

GEO accession no	**Publication**	**Tissue type**	**Number of samples**
Non-Neuronal Target matrix			
GSE3634	Abou-Sleymane *et al.*	Retina	2
GSE1816	Atul Butte *et a1.*	Retina	4
GSE3249	Cottet *et al.*	Retina	3
GSE4051	Akimoto *et al.*	Retina	5
GSE1835	Ruiz *et al.*	Retina	1
GSE5338	Cheng *et al.*	Retina	1
GSE3554	Steele *et al.*	Retina	1
GSE1999	Elizabeth Salomon *et al.*	Brain	2
GSE2867	Elizabeth Salomon *et al.*	Brain	4
GSE2869	Elizabeth Salomon *et al.*	Brain	1
GSE4675	Semeralul *et al.*	Brain	4
GSE4758	Elizabeth Salomon *et al.*	Brain	3
GSE4040	Elizabeth Salomon *et al.*	Brain	1
GSE6678	Elizabeth Salomon *et al.*	Brain	3
Non-Neuronal Target matrix			
GSE4067	Nilsson *et al.*	Muscle	2
GSE5304	Eric Hoffman *et al.*	Muscle	4
GSE1303	Eric Hoffman *et al.*	Lung	4
GSE6323	Edwards *et al.*	Muscle	2
GSE2198	Pederson *et al.*	Muscle	4
GSE6077	Cox *et al.*	Lung	1
GSE1471	Henry J. Kaminski *et al.*	Cardiac muscle	2
GSE1873	Li *et al.*	Liver	1
GSE3858	Cao *et al.*	Skeletal muscle	7
GSE1413	Majumder *et al.*	Prostate	2
GSE4711	Iguchi *et al.*	Testes	2
GSE2259	Denolet *et al.*	Testes	4
Reference matrix			
GSE6514	Mackiewicz *et al.*	Brain	16

Annexe 3: Array types available in RETINOBASE

Human

- ❑ Affymetrix GeneChip Human Genome U133 Plus 2.0 Array: ~ 47,000 probesets

Chicken

- ❑ Affymetrix GeneChip Chicken Genome Array: ~ 35,000 probesets

Mouse

- ❑ Affymetrix GeneChip Murine Genome U74 Set (Version2) consists of three probe arrays (MG-U74Av2, MG-U74Bv2, MG-U74Cv2): ~ 36,000
- ❑ Affymetrix GeneChip Murine Genome 430A 2.0 Array: ~ 22,600 probesets
- ❑ Affymetrix GeneChip Murine Genome 430 2.0 Array: ~ 45,000 probesets
- ❑ Cepko BMAP (Retinal) cDNA (spotted DNA array): ~ 12,500 cDNAs

Rat

- ❑ Affymetrix GeneChip Rat Genome U34 Array Set only (RG-U34A): ~ 8,000 probesets

Zebrafish

- ❑ Affymetrix GeneChip Zebrafish Genome Array: ~ 14,900 probesets

Drosophila

- ❑ Affymetrix GeneChip Drosophila Genome Array : ~ 13,500 probesets

Résumé

Le travail présenté dans ce manuscrit concerne différents aspects de l'analyse des données d'expression de gènes, qui englobe l'utilisation de méthodes statistiques et de systèmes de stockage et de visualisation, pour exploiter et extraire des informations pertinentes à partir de grands volumes de données. Durant ma thèse j'ai eu l'opportunité de travailler sur ces différents aspects, en contribuant en premier lieu aux tests de nouvelles approches de classification et de méta-analyses à travers la conception d'applications biologiques, puis dans le développement de RETINOBASE (http://alnitak.u-strasbg.fr/RetinoBase/), une base de données relationnelle qui permet le stockage et l'interrogation performante de données de transcriptomique et qui représente la partie majeure de mon travail.

Abstract

The work presented in this manuscript concerns different aspects of gene expression data analysis, encompassing statistical methods and storage and visualization systems used to exploit and mine pertinent information from large volumes of data. Overall, I had the opportunity during my thesis to work on these various aspects firstly, by contributing to the tests through the design of biological applications for new clustering and meta-analysis approaches developed in our laboratory and secondly, by the development of RETINOBASE (***http://alnitak.u-strasbg.fr/RetinoBase/.***) , a relational database for storage and efficient querying of transcriptomic data which represents my major project.

Keywords

Gene expression/Expression de gènes, Microarray, Meta-analysis/Méta-analyse, Dirichlet distribution/Distribution de Dirichlet, Clustering/ Classification, Relational DataBase Management Systems (RDBMS)/ Système de Gestion de Base de Données Relationnels (SGBD-R), Retina/ Rétine, RETINOBASE.

Laboratory

Laboratoire de Bioinformatique et de Génomique Intégratives (UMR7104), IGBMC, 1 rue Laurent Fries, 67404 Illkirch-Graffenstaden cedex, France.