Universidad Autónoma de Madrid

Facultad de Ciencias

Departamento de Biología Molecular

Functional profiling of genome-scale experiments

New approaches leading to a systemic analysis

Tesis Doctoral

Pablo Mínguez Paniagua

Madrid 2008

"Genoma: nuestra propia tragicomedia con cuatro únicos personajes, ácidos. El genoma es el teatro que cada uno de nosotros representa y ensaya en una eternidad de ¡bis! Las infinitas funciones de cada una de nuestras células crea el teatro génico de nuestro yo. El ADN es el micro-ordenador de nuestro macro-cerebro infra-utilizado por ser super-desconocido. En este escenario encerramos el Gran Teatro del Mundo y el de nuestra existencia".

Fernando Arrabal

Acknowledgments

En primer lugar quiero expresar mi más profundo agradecimiento a mi director de tesis, Ximo Dopazo, por darme la oportunidad de entrar en este mundo apasionante que es la ciencia, por introducirme en su fantástico grupo y por guiarme durante estos cuatro años de trabajo.

A la gente del labo, entre ellos muchos buenos amigos. Gracias a Jaime, Lucía y Álvaro por todos los buenos ratos que pasamos pero también por hacerme sentir arropado en los menos buenos. A Nacho, porque es muy buena gente. A Fátima, David y Stefan por toda su ayuda, esta tesis no la hice solo. También a Leo, Hernán (gracias por referenciarme), Fransuá, Pablo (the clever), Emidio, Eva, Paco, Toni, Joaquín, Josete, Ana, Marc, Patricia, Marina, Rafa, Davide. A los que ya no están por aquí, Jordi, Peio, Patricio y en definitiva a todos los que pasaron por bioinfo, seguro que llevándose un bonito recuerdo.

Gracias a Mamen, por los maravillosos años que compartimos, no estaría aquí sin tu ayuda.

A los Keyman, muy en especial a Imelda (gracias tronca por ser como eres) pero también a Mar y como no a Vicente. Gracias por todo, por cuidarme, apoyarme y alimentarme!, por tantos ratos divertidos sin mando a distancia junto al Titanic. Gracias muchas a los amigos de Valencia, algunos hasta compañeros del Flipe, gracias Seco y Rafa por los buenos ratos, los de agobios y los de risas, y en definitiva a toda la gente estupenda que anda por aquí, ha sido tan dificil aburrirse ... thanks for all tomorrows parties! Also to Guy, thanks for your no-classes and specially for our liquid dinners, best of luck mate!

A mis amigos, muy en especial a Antonio por cuidarme siempre y compartirse conmigo, un lujo tenerte cerca. A Rafa y a Raúl, con ellos compartí momentos muy especiales de mi vida. También a Angelo, por todo su apoyo y todos los buenos momentos que pasamos en esas tierras de cambios y aventuras. Y en definitiva, a la gente que en algún momento se ha cruzado en mi camino para hacerme sonreír, aquellos con los que compartí días de estudio y a los que no me dejaron estudiar, a los que siguen por aquí y a los que algún día volveré a ver.

Muchas gracias también a los miembros del tribunal.

Special thanks to The Ramones, they introduced me into all this, biology does not own you much guys but I certainly do. Al Magik, el Turmix, Aguacates, el Asesino y la Bounty, y a esta Valencia en la que he vivido mucho. A la Guinness, los Echo, las sopas de ajo, los resopones, Panero, Allen, el Rose Tavern y el Garden House, las tapas de mi pueblo y a tanta buena gente con la que he compartido todo esto.

A Pilar, gracias mil por mantenerme de buen humor durante la escritura, por escucharme y apoyarme. Por nuestros gin&tonics compartidos y todos los que deseo que vengan, eres fantástica.

A mi familia, mis abuelos, mis tíos y mis primos, sois todos maravillosos. A mi hermana, a la que quiero muchísimo y a David (mi mejor cuñado), gracias por toda vuestra ayuda y comprensión.

Y finalmente gracias a mis padres, a quienes dedico esta tesis, porque transmitís mucho amor, todo lo bueno que sé lo aprendí de vosotros.

Contents

Part I

Introduction

Chapter 1

Functional Genomics and Systems Biology

Functional genomics is the field of molecular biology that attempts to describe cell behaviour from the data produced by genome-scale experiments. Historically, genes and proteins (their functionally active form) have been defined as the functional units in the cell. Focusing on this assumption, molecular biology reductionist approach has given excellent advances in the basic understanding of living organisms by the identification and description of the components responsible for particular processes in their cells and tissues. Despite this success, many fundamental biological questions remain unanswered, mainly because there are very few processes that can be explained by the action of a single protein. On the contrary, the units of activity involved in cellular processes seem to be modules composed by several interacting molecules (Hartwell *et al.*, 1999; Barabasi and Oltvai, 2004). This fact represents a limitation for the classical molecular biology techniques based on the description of the action of one or few genes (e.g. northern blot) or proteins (e.g. western blot) at a time. In recent years, several high-throughput experimental methodologies has arisen with the goal of investigating the action of, ideally all, but in practice thousands of genes or proteins simultaneously. Microarray experiments (Schena *et al.*, 1995) which can probably be considered the milestone in this type of techniques, report the expression level of a great proportion of the transcripts in the cell. Other techniques that may have the same application among others are Serial Analysis of Gene Expression (SAGE, Velculescu *et al.*, 1995) and Expressed Sequence Tags (ESTs, Adams *et al.*, 1991). Very recently, next generation sequencing technologies are starting to be applied to the study of the transcriptome. There are other high-throughput techniques that focus on the post-translational events, two-dimensional gel electrophoresis that explore protein abundance or yeast two hybrid assays that extract protein-protein interactions are two relevant examples.

The set of all transcripts (messenger RNA, mRNA) produced in a cell in a particular condition is generally named as the transcriptome. Contrarily to

the static property of the genome (hereditary information encoded in DNA), the transcriptome is a dynamic entity whose elements vary their presence and quantity depending on the cell nature and state. The motivation of this thesis was to develop methodologies that permit to extract significant modules of action from the transcriptome.

Hereby, the cell seems to have a complex machinery that cannot be summarized as the action of their genes isolated. To understand the intricate network of interactions among all types of cell components (genes, proteins, metabolites, etc.) we need to know both, the features of the elements separately and the consequences of their cooperative behaviour. This new emergent property cannot be studied under a flat perspective but under a holistic view of the cell. Systems biology is the field of science that is being used to undertake this new approach, it attempts to describe cell behaviour in terms of the quantification of the interaction among all its individual components.

Systems biology assays agree on the bases of the characterization of a cell state in systems terms (Kitano, 2002; Bruggeman and Westerhoff, 2006), setting as compulsory the description of four elements:

- The structure of the system, including the elements, the interaction networks and the pathways that conform the system as well as the mechanisms that translate such structure into a phenotype.
- The system dynamics, that is, how the elements and their relationships evolve over time under certain conditions.
- The control method that tries to minimize perturbations in the system.
- A design method able to model the system and predict its features under determined conditions.

A functional profiling analysis of a high-throughput experiment is the process of describing at molecular level the functionalities responsible for a particular phenotype.

Systems biology can be viewed as the framework in which functional genomics experiments are analysed via functional profiling with an integration of "omic" data (see next section for a definition). The aim is to get as much as possible knowledge about the system under study to be able to build a model capable of predicting its behaviour. The first three points of the systems biology requirements can be addressed by different functional genomics experiments. Thus, the more omic data we integrate, the better performance we will obtain in the functional profiling analysis. The last point needs the participation of mathematical modelling algorithms which are beyond the scope of this thesis. This thesis is dedicated to the part where functional genomics and functional profiling may participate, and it is conceived from the beginning under a systems biology perspective.

Chapter 2

Omics and its integration

The necessity of obtaining a complete knowledge of the cell elements and their functional relationships in order to build a model that explain its behaviour has promoted the proliferation of novel techniques that screen the population of several types of biological molecules such us proteins, mRNAs or metabolites and the set of actions performed by them: protein-protein interactions, the quantification of their fluxes, etc. That kind of data is generally included under the neologisms -omic and -ome that refers to biological studies such as proteomics, genomics and the data they generate (proteome, genome).

Omics and omes have arisen as a revolution in biology. In the early years, the bottleneck of molecular biology was the production of new data (gene discovery, interactions discovery, etc.). Typically we had a lot of information from a few genes. In this new era we are in the opposite situation, there is a huge amount of data and little first hand knowledge about it.

The first ome term was genome, a word adapted by Hans Winkler in 1920 as a fusion of gen(e) and (chromos)ome to refer to the set of genes in all the chromosomes. Nowadays it includes all the hereditary information, coding and non-coding sequences. Although -ome and -omic are not known roots in any language, they have been adopted as neologisms to define the complete compilation of elements of a set and the field that study them respectively. Thus, an explosion of omics terms have recently appeared in the literature with different success, table 2.1 shows some of the omes that has been used.

Omics data is mostly generated by high-throughput experiments characterised for producing a huge amount of data. The process of curation of this data is a hot topic in modern molecular biology. There is a clear necessity of controlling false positives and negatives. Below in this thesis we will discuss some of the methodologies for curation of interactomics data. The storage of this data is also a new problem for molecular biologists. The necessity of well designed relational databases that permit an easy and quick query system to the data has been one of the main tasks for bioinformaticians in last ten years.

To be able to transform this huge amount of data into information we need to

fulfil at least three requirements. Two of them has been mentioned before: curation and manageability. The third one is its integration. Indeed, the information at only a unique level (genome or proteome for example) by itself cannot fully explain the behaviour of any particular biological system. We may cite as an example the lack of correlation between protein and mRNA abundance in yeast (Gygi *et al.*, 1999) and human liver (Anderson and Seilhamer, 1997). In recent years several post-transcriptional regulation agents such as miRNA (Lee *et al.*, 1993) and siRNA (Hamilton & Baulcombe, 1999) have been discovered.

This thesis came up from the beginning with the aim of developing methodologies capable of integrating as much sources of information as possible under the prism of systems biology. To continue with the omics fever within the computational biology environment we may say that this thesis is also intended to contribute to integromics, yet another omic field defined as the integration of several omics. Integromics will be a crucial step forward for the translation of data into information.

Omic term	Description	Google search	Pubmed entries	First year in Pubmed
Genome	The full complement of genetic information both coding and non coding in the organism	60,400,000	637,127	1943
Proteome	The protein-coding regions of the genome	9,100,000	11,370	1995
Transcriptome	The population of mRNA transcripts in the cell, weighted by their expression levels	2,250,000	37,601	1997
Phenome	Qualitative identification of the form and function derived from genes, but lacking a quantitative, integrative definition	1,720,000	96	1995
Interactome	List of interactions between all macromolecules in a cell	142	277	1999
Metabolome	The quantitative complement of all the small molecules present in a cell in a specific physiological state	111	431	1998
Orfeome	The sum total of open reading frames in the genome, without regard to whether or not they code; a subset of this is the proteome	107	42	2002
Kinome	The population of protein kinases in the genome	53,300	117	2002
Physiome	Quantitative description of the physiological dynamics or functions of the whole organism	46,700	64	1997
Secretome	The population of gene products that are secreted from the cell	43,200	158	2000
Glycome	The population of carbohydrate molecules in the cell	15,500	69	1999

Omic term	Description	Google search	Pubmed entries	First year in Pubmed
Fluxome	The population of proteins weighted by their fluxes	9.,670	20	1999
Regulome	Genome-wide regulatory network of the cell	9,090	13	2004
Morphome	The quantitative description of anatomical structure, biochemical and chemical composition of an intact organism, including its genome, proteome, cell, tissue and organ structures	6,560	3	1996
Lipidome	Compilation of lipids in a cell	5,420	34	2006
Localizome	The localization of various proteins, both in terms of cell type and subcellular compartments	3,260	3	2001
Translatome	The population of mRNA transcripts in the cell, weighted by their expression levels	1,970	3	2001
Phylome	Complete collection of all gene phylogenesis in a genome	1,660	4	2001
Cellome	The entire complement of molecules and their interactions within a cell	1,540	28	2002
Transportome	The population of the gene products that are transported; this includes the secretome	1,290	4	2004
Functome	The population of gene products classified by their functions	598	1	2001
Ribonome	The population of RNA-coding regions of the genome	309	1	2002
unknome	Genes of unkown function	216	-	-
Foldome	The population of gene products classified by their tertiary structure	167	-	-
Operome	The characterization of proteins with unkown biological function	141	-	-

Table 2.1: List of some of the ome terms more used in the literature. The table contains a small definition as well as the number of entries in a google search and in a Pubmed search as a measure of their usage. Table updated by the author on 16th of July, 2008, the original table taken from `http://bioinfo.mbb.yale.edu/what-is-it/omes/omes.html`.

Chapter 3

Functional profiling

The functional profiling of high-throughput experiments requires basically of two elements: sources of gene and protein annotation and methodologies capable to extract the important cellular processes that define the cell behaviour in a particular state. In the next subsections we will give a short description of the sources of annotation used in this thesis followed by a small review on the methods available for functional profiling.

3.1 The Biological sequences' annotation

We need to have a definition of modules of action to be able to search for them within the data reported by the high-throughput experiments. This definition comes from the sources of annotation. The assumption of these methods is that functionally related genes tend to co-express (Stuart *et al.*, 2003; Lee *et al.*, 2004). They are, together with the results coming from the high-throughput experiments, the two input parameters of the methodologies developed to perform functional profiling. In fact, the methodologies are developed as they can extract the maximum information from the results of the experiments taking into account the special nature of the annotation. A good knowledge on how the different types of annotations are structured, their degree of curation and how is the procedure of the annotation process is a fundamental requisite before approaching the development of a methodology for functional profiling.

Generalizing, in this thesis we have used three types of annotation:

- **Discrete labels**, such as Gene Ontology terms, KEGG pathways and BioCarta pathways. Their association to the gene is in *have it or not* terms. They may have a flat internal structure as KEGGs and BioCarta or be a structured vocabulary as GO terms.
- **Continuous labels**, associated to the genes or proteins through a value. In this thesis we will report methodologies using two different annotations in

this category: words associated to genes extracted from scientific literature using text-mining techniques and genes associated to different tissues and histologies (phenotype) through an expression measurement.

- **Discrete labels with a supra-structure**, this is the case of protein-protein interaction data where every protein is associated to other proteins conforming a network where the nodes are the proteins and the edges are the interaction events. The network itself have intrinsic features that cannot be described as the sum of its parts.

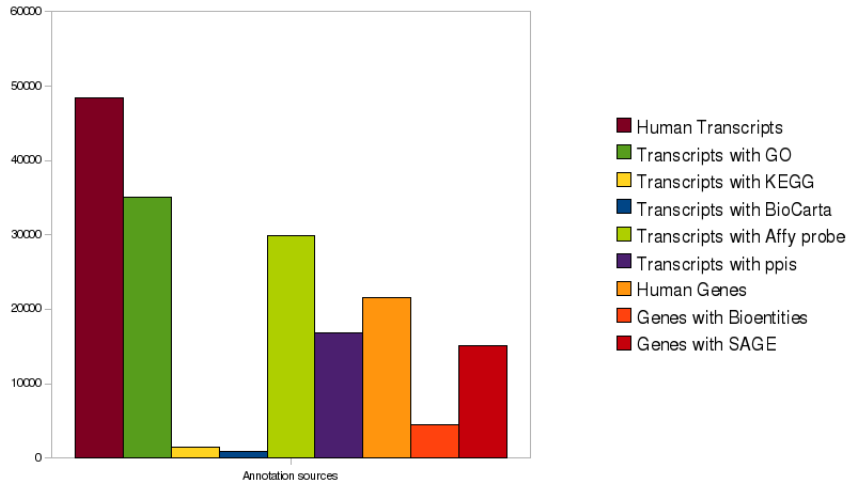


Figure 3.1: Annotation sources coverage in human. The histogram represents the number of transcripts and genes in ensembl database (v49) and the coverage of the sources of annotation used for the functional profiling of experiments in this thesis, GOs, KEGGs, phenotype (Affymetrix probes) and BioCarta referenced to transcripts and bioentities extracted from literature and phenotype (SAGE tags) referenced to genes.

The coverage of the sources of annotation in human genome and proteome is shown in figure 3.1.

3.1.1 Discrete labels

3.1.1.1 Gene Ontology

Undoubtedly, the most used annotation source in functional genomics is the Gene Ontology (GO), proposed by the Gene Ontology consortium (`http://www.geneontology.org`). It provides a controlled vocabulary for the description of molecular function, biological process and cellular component of gene products (Ashburner *et al.*, 2000). Nowadays, GO terms can be considered as the most standard gene product annotation. GO is being used by most of the gene and protein databases, facilitating enormously the querying to the end users and the management of annotation through computers by computational biologists. GO

annotation is hierarchical, that means that a sequence is annotated with different grades of specificity. Every term has a unique numerical identifier of the form *GO:xxxxxx* and a more meaningful name (e.g. *GO:0045777* refers to *positive regulation of blood pressure*).

Terms are structured in form of a Directed Acyclic Graph (DAG) that is similar to a tree topology where each node is connected to other nodes in several types of relationships:

- *is_a* relationship means that a child term is an instance of the parent (e.g. *chloroplast envelop GO:0009941* is a *membrane GO:0016020*).
- *part_of* relationship refers that the child is a component of the parent (e.g. *inner membrane GO:0019866* and *outer membrane GO:0019867* are components of *membrane GO:0016020*).
- *regulates*, *positively_regulates* and *negatively_regulates* are relationships that describe interactions where a GO term modulates the occurrence or the value of another GO term.

The difference between a DAG and a tree is that in a DAG a term may have more than one parent. The deeper a node is in the hierarchy, the more detailed is the description of the term.

The DAG starts with an universal root GO term named *all:all* located at level 0. As children (level 1) it has three not connected terms that represent three different ontologies:

- *Molecular function* (*GO:0008639*, MF or F), defined as the actual functionality of a gene product at a molecular level, in other words, its activity within the cellular machinery.
- *Biological process* (*GO:0008150*, BP or P), that represents a collection of molecular events with a defined beginning and end. Within this category the gene products are annotated according to the processes in which they are involved.
- *Cellular component* (*GO:0005575*, CC or C) refers to the part of a cell or its extracellular environment in which a gene product is located.

3.1.1.2 KEGG pathways

The Kyoto Encyclopedia of Genes and Genomes (KEGG, `http://www.genome.jp/kegg/`) provides a variety of databases related to molecular biology and biomedicine. Some of them deal with genes, proteins, chemical reactions, compounds, drugs and pathways. The KEGG Pathway Database (Kanehisa *et al.*, 2004) is a well known repository for curated biochemical pathways. The genes are annotated to participate in any of the reactions belonging to a specific pathway.

The pathways are classified into 6 different categories: *metabolism*, *genetic information processing*, *environmental information processing*, *cellular processes*, *human diseases* and *drug development*. They are in turn subdivided into more specific subcategories (e.g. *genetic information processing* has as subcategories: *transcription*, *translation*, *folding*, *sorting and degradation* and *replication and repair*). The coverage of KEGG pathways is not as extense as in the case of GO terms but its definitions are more reliable because the annotation is always manually curated.

3.1.1.3 BioCarta

BioCarta (`http://www.biocarta.com`) is another curated repository that annotates genes in terms of their participation in molecular pathways. As KEGG, it has a quite simple classification. The pathways subcategories are *adhesion*, *apoptosis*, *cell activation*, *cell cycle regulation*, *cytokine/chemokines*, *developmental biology*, *hematopoiesis*, *inmunology*, *metabolism* and *neuroscience*.

3.1.2 Continuous labels

3.1.2.1 Bioentities extracted by text-mining techniques

Indeed, the annotation of biological sequences is the principal input component for functional profiling methodologies. Although there are a lot more initiatives than the mentioned above that try to achieve a curated and complete annotation of genes in different fields of the scientific knowledge, we are still far away from the goal of a complete coverage of the genomes. Thus, there is a clear necessity for standard and curated annotation that provides bioinformatics tools with reliable annotations to be able to generate other kind of knowledge. Nevertheless, the scientific community has been generating high quality information over hundreds of years. In fact, the biggest encyclopaedia about functional genomics is not in the databases but in scientific journals and books as free text (also in tables, pictures and graphics). The extraction of this information and its storage in the new developed databases and functional genomics resources is of crucial importance, although not a trivial task.

The increasing interest in developing computational methods to extract high quality and manageable information from free text is a common goal in many fields from marketing to security. In the case of biomedicine, a large collection of abstracts and articles are electronically available through the Medical Literature Analysis and Retrieval System Online (MEDLINE), a literature database compiled by the National Library of Medicine (NLM) from 1964. MEDLINE is freely available on the internet through the service PubMed, part of the Entrez retrieval system for biological knowledge (Schuler *et al.*, 1996). This constitutes an excellent raw material for the application of the called natural language processing (NLP) techniques that deal with natural language and free text analysis. NPL has two main applications that are worth to be mentioned:

- The information retrieval (IR) techniques, that focus on extracting textual information from a collection of documents. This is what PubMed does to MEDLINE database and Google to the internet.
- The information extraction (IE) techniques, that tries to automatically extract structured information such as patterns or relationships among words from free text.

IE methodologies are being used extensively for the automatically annotation of genes and proteins. In this thesis, we have used annotation data generated by the AlmaKnowledgeServer software (`http://www.bioalma.com/aks2/`), a successful instance of this kind of approaches, to the functional profiling of genome scale experiments.

3.1.2.2 Gene expression in different tissues (phenotype data)

The genes can also be characterized by their transcriptomic information, that is, the type of cells, developmental stage and histology status in which they are transcribed and at what level. In the case of a transcriptome analysis, for instance a microarray experiment exploring the differences between a cancer and a normal tissue, we might want to know whether the genes over-expressed in the cancer sample are specific to that phenotype, appear in many types of cancer, are housekeeping genes or are associated to other kind of dysfunction or developmental stage not related to the cancer at all. This can be done by reporting the levels of expression of the genes over different conditions and cell types making use of the wide collection of transcriptomic experiments available in the public databases.

In this thesis we have used two types of transcriptomic experiments to annotate genes and proteins, Serial Analysis of Gene Expression (SAGE) and DNA microarrays.

3.1.2.2.1 Serial Analysis of Gene Expression (SAGE) SAGE is a well known transcriptome exploration technique that set up its basics in two principles:

- A few nucleotides may theoretically identify uniquely a transcript. Indeed, if we consider 4 bases and 9 nucleotides, the number of possible combinations are $4^9 = 262{,}144$, which are more than enough to identify the transcripts produces by the human genome.
- There are restriction enzymes that have the capability of cutting off a sequence in a determined position, e.g. NlaII cut in the pattern 5'-CATG-3' closer to the polyA chain.

The generation of a SAGE library has the following steps:

- Isolate mRNA from an input sample (e.g. a tumour).
- Extract a small chunk of sequence from a defined position of each mRNA molecule (restriction enzyme).
- Link these small pieces of sequence together to form a long chain (concatemer).
- Clone these chains into a vector which can be taken up by a bacteria.
- Sequence these chains using high-throughput DNA sequencers.
- Process this data with a computer to count the small sequence tags.

3.1.2.2.2 DNA Microarrays Without any doubt, microarrays are the most extensively used high-throughput technique to determine the transcriptome of a cell nowadays. There are mainly two categories of microarrays depending on the type of probes they use to detect gene expression: cDNA or oligonucleotides.

In cDNA microarrays the probes are molecules of cDNA. Their length vary between 500 and 5,000 bases and are also called two colours microarrays because the expression is measured as a ratio of the expression of two samples labelled with different fluorophores (e.g. Cy3 and Cy5) that have competitive hybridization for the probes. The second type of microarrays has probes made of oligonucleotides of variable length, typically from 25 to 70 bases. Possibly the major producer of this type of chips is Affymetrix. Affymetrix chips have pairs of probes to detect gene expression, every pair of probes has a perfect match probe and a mismatch probe which controls the unspecific hybridization so no control sample is needed. They are generally called one colour arrays.

3.1.3 Discrete labels with a supra-structure

3.1.3.1 Protein-protein interactions

Protein-protein interactions (ppis) play a central role at almost every level of cell activity: they are involved in the structure of organelles (structural proteins), transport machinery (nuclear pore importins), response to stimulus (signalling cascades), regulation of gene expression (transcription factors), protein modification (kinases) among many other processes. The production and the proper use of this type of information is of crucial importance in order to understand cell behaviour. The available ppi data has increased enormously in the last few years with the emergence of high-throughput techniques that can report thousands of ppis in a short time span. The most used techniques in this field are: yeast two hybrid (y2h), tandem affinity purification (TAP) and high-throughput mass spectrometry techniques (MS). Reviews on these and related methodologies can be found in Drewes and Bouwmeester (2003), Cho *et al.* (2003), Falk *et al.* (2007) and Berggard *et al.* (2007).

The reliability of this data is not exempt of controversy. Studies comparing resulting data from several experiments demonstrate that the overlap between them is not as extensive as desirable. This can be because the methods do not reach the saturation point (Bader & Hogue, 2002) or due to the lack of accuracy and coverage on some of them (von Mering *et al.*, 2002). In spite of this, there are arguments in favour; each experiment may cover only 3-9% of the total interactome, so limited overlap should be expected (Han *et al.*, 2005). False positives are also a problem: in y2h these represent up to 50% of the total data (Ito *et al.*, 2001; Mrowka *et al.*, 2001). Moreover, there is a bias in the functional categories of the ppis each technique detects, e.g. y2h fails in detecting proteins involved in translation (von Mering *et al.*, 2002).

The ultimate objective of all these techniques is to generate a complete map of all possible ppis that can potentially occur in the cell, this map is commonly known as the interactome. And, beyond discussions about accuracy and coverage of this kind of experiments, the relevance of ppis in the cellular machinery has fostered an unprecedented interest in the exploration of the interactome of model organisms such as *Saccharomyces cerevisiae* (Uetz *et al.*, 2000; Ito *et al.*, 2001), *Drosophila melanogaster* (Gio *et al.*, 2003; Formstecher *et al.*, 2005), *Caenorhabditis elegans* (Li *et al.*, 2004) or human (Stelzl *et al.*, 2005, Rual *et al.*, 2005), just to cite a few examples.

In yeast, a high-quality literature curated set of ppis free from false positives and representing probably the complete interactome (Reguly *et al.*, 2006) is available. However, in the case of human, the situation is far away from this degree of detail. The estimated size of the human interactome is of 650,000 ppis (Stumpf *et al.*, 2008). None of the public databases contain more than 10% of this number of ppis, and a compilation of all the known ppis would only cover about 10% of the interactions.

The interactome so obtained is an abstract scaffold that does not provide information about particular conditions, cell developmental stage or cell type in which a particular ppi occurs (if any). To infer a case-specific interactome it is necessary to integrate other types of data that provide information that allows inferring the active ppis at a particular condition.

3.1.3.2 Bases for detecting ppis and its annotation

In this new era of massive production of biological data, an important challenge is its storage in a standardised format with appropriate annotation that facilitates performing queries as simple as possible to extract relevant information. Several datasets coming from high-throughput technologies such as biological sequences or microarray experiments have developed structured formats to submit the data to the databases with ontology based vocabulary. Learning from those experiences, the Proteomic Standards Initiative (PSI) of the Human Proteome Organization (HUPO) has established a Molecular Interaction (MI) group to develop a standard format to interchange information called PSI-MI (Hermjakob *et al.*, 2004). In here we report on the main categories MI has to classify the

experimental detection methods:

- *biophysical*: The application of physical principles and methods to biological experiments.
- *protein complementary assay*: The function of numerous proteins (enzymes, transcription factors, and others) can be rationally dissected into two fragments that fold autonomously but cannot complement to reconstitute the complex function, unless they are located in close proximity. In a two hybrid experiment, restoration of the activity by complementation of the two fragments when expressed as fusion with two polypeptides is taken as an evidence that the two polypeptides interact together.
- *genetic interference*: This term refers to methods that aim at interfering with the activity of a specific gene by altering the gene regulatory or coding sequences. This goal can be achieved either by a classical genetic approach (random mutagenesis followed by phenotype characterization and genetic mapping) or by a reverse genetics approach where a gene of interest is modified by directed mutagenesis.
- *post translational interference*: This term refers to methods designed to interfere with gene expression at post-transcriptional level rather than with the gene itself.
- *biochemical*: The application of chemical principles and methods to biological experiments.
- *imaging techniques*: Methods that provide images of molecules at various resolution depending on the technology used.

3.1.3.3 Ppi data resources (Databases)

At the time of writing this thesis, there is not a common repository that stores all the ppis. Contrarily to other genomic data such as sequences, microarrays, protein structures, etc., ppi data are spread through several databases, among which a small overlap exists. Moreover, there are differences in the type and depth of ppi annotations among the databases. The major repositories are the Human Protein Reference Database (HPRD, Peri *et al.*, 2003), IntAct (Kerrien *et al.*, 2006), the Bimolecular Interaction Network Database (BIND, Bader *et al.*, 2003), the Database of Interacting Proteins (DIP, Salwinski *et al.*, 2004), BioGRID (Breitkreutz *et al.*, 2008) and the Molecular INTeractions database (MINT, Chatr-aryamontri *et al.*, 2006). Therefore, it is not a trivial task for the end user to obtain a reasonably complete and curated set of ppis to work with. Several methodologies have been proposed to solve this problem (see next section for a small revision on them). In the Materials and Methods section 14.2 we propose a novel method for this purpose.

Reviews on the resources dedicated to store and annotate ppis can be found in Xenarios and Eisenberg (2001) and Mathivanan *et al.* (2006).

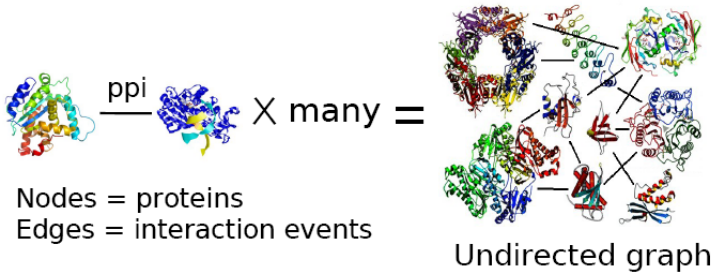


Figure 3.2: From ppis to the interactome. Several pairwise protein-protein interactions can be represented as an undirected graph.

3.1.3.4 Ppis as networks (The supra-structure)

Ppis are defined as pairwise relationships between two proteins. Taken all together, they represent a network where the nodes are the proteins and the edges the interaction events (see figure 3.2). Apart from the elements of the network (nodes and edges), the topology of the networks is also of crucial importance when trying to understand their role in a cellular process (Yeger-Lotem *et al.*, 2004).

Graph theory has helped biology to study these networks and has established the basis for their description. The first discovery was that biological networks are scale-free networks (Barabasi & Albert, 1999; Barabasi & Bonabeu, 2003) instead of random networks. Scale-free networks are defined by a connections degree, number of connections of a node, distribution that approximates to a power law $P(k) = k^{-\gamma}$, being $\gamma < 3$. This indicates that the network has a low number of highly connected nodes, called hubs. In other words, there are a few proteins (the hubs) which connect much of the whole network. Indeed, identifying hubs is a hot topic in functional analyses (Batada *et al.*, 2006; He & Zhang, 2006; Sporns *et al.*, 2007).

Apart from connections degree, which identifies the presence of hubs, there are other network parameters that help to describe properties of these systems (Barabasi & Oltvai, 2004). In the Materials and Methods section 6.4.1 we report all the network features used in this thesis to describe complete interactomes, sub-networks as well as other kind of networks.

3.2 Methodologies for functional profiling

3.2.1 Functional enrichment methods

A conventional approach to the interpretation of genome-scale data is usually performed in two steps: in a first step genes of interest are selected (because they co-express in a cluster or they are significantly over- or under-expressed when

two classes of experiments are compared, etc.) and then the enrichment of any type of biologically relevant label in these genes is compared to the corresponding distribution of the label in the background (typically the rest of genes in the genome or in the experiment). There are different available tools, such as FatiGO (Al-Shahrour *et al.*, 2004) and others (Zeeberg *et al.*, 2003; Khatri and Draghici, 2005), that use different functionally relevant labels such as GO terms (Ashburner *et al.*, 2000), KEGG pathways (Kanehisa *et al.*, 2004), etc. A revision of functional enrichment methods can be found in Dopazo (2006).

In this thesis we have developed methods that follow this approach for non-conventional annotation sources like ppi data, bioentities extracted from the literature and expression levels in tissues.

3.2.1.1 The multiple testing problem

Much caution should be adopted when dealing with a large set of data because of the high occurrence of spurious associations (Ge *et al.* 2003). Addressing multiple testing properly is a rather complex problem. Many of the conventional correction methods (e.g. Bonferroni or Sidak) are based on the consideration that a p value should be adjusted by multiplying a reasonable significant threshold (e.g. $p< 0.05$) for the number of tests performed to obtain a new threshold. Whenever many thousands of tests are performed the original assumption risks to be too conservative. A better strategy to estimate p values is provided by another family of methods that allow less conservative adjustments are the family wise error rate (FWER), that controls the probability that one or more of the rejected hypotheses (GO terms whose differences cannot be attributed to chance) is true (that is, a false positive). The minP step-down method (Westfall and Young, 1993), a permutation-based algorithm, provides a strong control (e.g. under any mix of false and true null hypothesis) of the FWER. Approaches that control the FWER can be used in this context although they are dependent on the number of hypotheses tested and tend to be too conservative for a high number of simultaneous tests. Aside from a few cases in which FWER control could be necessary, the multiple testing problem in functional assignation does not require protection against even a single false positive. In this case, the drastic loss of power involved in such protection is not justified. It would be more appropriate to control the proportion of errors among the identified GO terms whose differences among groups of genes cannot be attributed to chance instead. The expectation of this proportion is the False Discovery Rate (FDR). Different procedures offer strong control of the FDR under independence and some specific types of positive dependence of the tests statistics (Benjamini and Hochberg, 1995), or under arbitrary dependency of test statistics (Westfall and Young, 1993).

3.2.2 Gene set enrichment analyses (Threshold-free methods)

Although widely accepted and with considerably good results in its application, the functional enrichment methods present some inconveniences associated to the

imposition of a threshold in the case of differential expression analyses. In this kind of studies, the "important" genes are selected using exclusively the expression data by applying a threshold in the p-values. The biological information is introduced *a posteriori* and it is then not used for the selection of the genes. Besides, high-throughput techniques are still particularly noisy, thus the imposed threshold must leave out a big number of false negatives to keep a low rate of false positives, ending with an incomplete list of genes to be analysed.

Under a systems biology perspective, this way of understanding the molecular basis of a genome-scale experiment is far away from being efficient. Methods inspired in systems biology focus on collective properties of genes. Functionally related genes need to carry out their roles simultaneously in the cell and, consequently, it is expectable from them to display a coordinated expression. Actually, it is a long recognized fact that genes with similar overall expression share often similar functions (Lee *et al.*, 2004; Eisen *et al.*, 1998; Wolfe *et al.*, 2005). This observation is consistent with the hypothesis of modularly-behaving gene programs, where sets of genes are activated in a coordinated way to carry out functions. Under this scenario, a different class of hypothesis, not based on genes but on blocks of functionally related genes, can be tested. Thus, lists of genes ranked by any biological criteria (e.g. differential expression when comparing cases and healthy controls, etc.) can be used to directly search for the distribution of blocks of functionally related genes across the list without imposing any arbitrary threshold. Any macroscopic observation that causes this ranked list of genes will be the consequence of cooperative action of genes that are part of functional classes, pathways, etc. Each functional class "responsible" for the macroscopic observation will, consequently, be found in the extremes of the ranking with highest probability.

There are different methods which have been proposed for this purpose such as the GSEA (Mootha *et al.*, 2003; Subramanian *et al.*, 2005) or the SAFE (Barry *et al.*, 2005) method that use a non-parametrical version of a Kolmogorov-Smirnov test. Other strategies are also possible, such as the direct analysis of functional terms weighted with experimental data (Smid *et al.*, 2004) or model-based methods (Goeman *et al.*, 2004). With similar accuracy although conceptually simpler and quicker methods have also been proposed such as the parametrical counterpart of the GSEA, the PAGE (Kim *et al.*, 2005) or the segmentation test, Fatiscan (Al-Shahrour *et al.*, 2005). Revisions on Gene Set methods can be found in Goeman and Bühlmann (2007) and Dopazo (2008).

FatiScan can deal with any kind of discrete labels, GO terms, KEGG pathways, etc. As seen before there are other kind of annotations that associate the labels to genes through a value. In this thesis we have developed a gene set enrichment method based on FatiScan that can deal with continuous labels, its first application has been to perform functional profiling using bioentities extracted from the literature by text mining techniques. It has been implemented with the name of MarmiteScan as a module in the Babelomics suite.

3.2.3 Functional enrichment methods using ppi data

The development of the methodologies for functional profiling is directly guided by the nature of the annotation that it is going to be used. We will show in Materials and Methods section 6.2 how same approaches vary in their basis when changing the annotation from discrete to continuous labels. In this section we will review the boinformatics tools to manage and visualize ppi data to finally explain the available methodologies for the analysis of genome-scale experiments using protein-protein interactions as annotation source.

The introduction of ppi data into functional genomics requires of new algorithms due to the particular nature of this kind of annotation. The way in which ppis are defined do not conform discrete classes (as GOs, KEGGs, etc.) but they are internally structured as networks where proteins have different roles according to their position in the network and its global shape. Therefore different methods are necessary to build these classes from expression and interactomic data.

A module in a network is a sub-network with an internal connectivity higher than its connectivity to other modules. Many attempts have been made to explore the interactome seeking for modules of action, most of them based on the application of clustering methods to weighted matrices, Pereira-Leal *et al.* (2004) proposed the number of experiments that support the interaction as index to fill the matrix, Rives and Galitski (2003) used the shortest paths among pairs of nodes to measure de relationship between nodes. There are also other approaches based, for instance, on the topological features of the network such as the betweenness (Girvan and Newman, 2002; Wilkinson and Huberman, 2004). Central nodes, with a high betweenness, may define the boundaries of the sub-networks because that means that many shortest paths pass through them and the action of removing them from the network would lead to the disconnection of some sub-networks.

Modules obtained by this methodologies may be enriched in proteins with related biological functionalities, shown by its significant enrichment in GO terms (Luo *et al.*, 2007) or by its co-occurrence within the literature (Wilkinson and Huberman, 2004). Indeed, it has also been shown that there are sub-networks associated to diseases (Badano and Katsanis, 2002; Brunner and van Driel, 2004; Gandhi *et al.*, 2006). Gandhi *et al.* (2006) found in the analysis of the human interactome that proteins encoded by genes mutated in inherited genetic disorders are likely to interact with proteins known to cause similar disorders.

When seeking for modules of action (sub-networks) within lists of genes or proteins, the typical functional genomics input, the list has not to be considered any longer as a mere collection of more or less important nodes but as a potential unit of functionality in cell activity, just as Gene Ontology terms or KEGG pathways are defined. It is not enough to just obtain the interactions associated to each of them and explore the function of the interacting proteins. Thereby, we need to build system specific functional modules within the set of proteins in order to assign a common functionality to the list, in other words, it is essential for the biological interpretation of a gene or protein set using ppi data to seek for the sub-network that they might form, this is, the module that has been

activated.

A common approach to figure out this sub-network is to calculate the called Minimal Connected Network (MCN). The MCN is the minimal network that connects a set of nodes. See Materials and Methods section 6.4.3 for a detailed explanation.

Chapter 4

The working environment

The methodologies presented in this thesis have been implemented as web tools and are part of two interconnected suite of programs, GEPAS (Tarraga *et al.*, 2008) for analysis of microarray experiments and Babelomics (Al-Shahrour *et al.*, 2008) for functional profiling of transcriptomics, proteomics and genomics experiments. This fact makes the resources freely available for the scientific community. Nowadays, many studies in molecular biology include a microarray experiment to set up the scenario in which the particular phenotype is occurring. In silico experiments are becoming a common practice in molecular biology and scientists do not stop in the analysis of their important genes but they try to see the case under study from a systems biology perspective. The goal of both GEPAS and Babelomics is to help biologists to use appropriate contrasted methods to give biological meaning to the messy data coming from high-throughput experiments. In figure 4.1 a road map of the GEPAS and Babelomics possibilities is shown.

The introduction in GEPAS and Babelomics of the methods presented in this thesis represent an effort for integrating new sources of annotation into the functional profiling of high-throughput experiments.

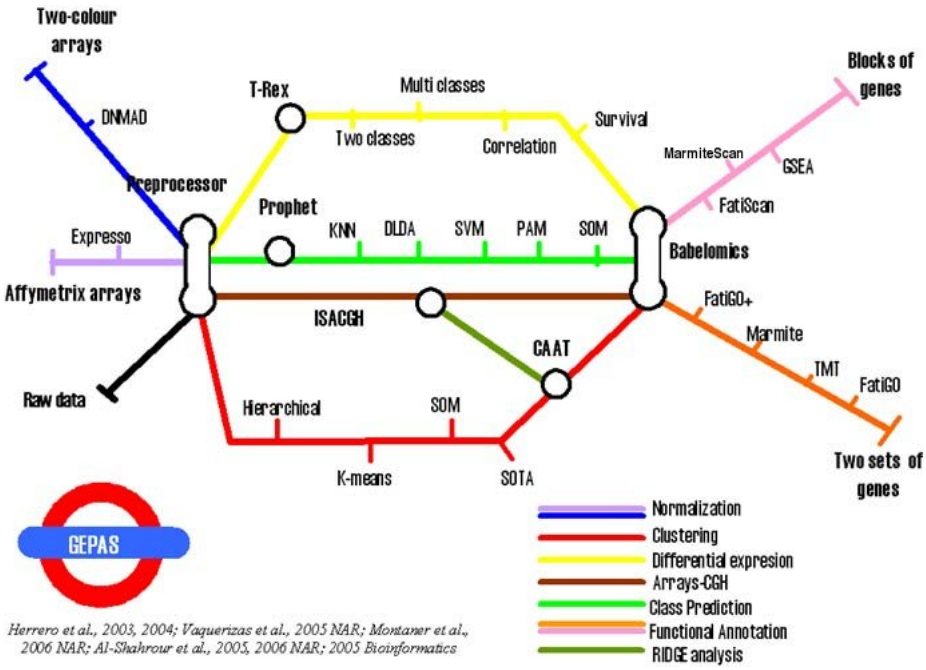


Figure 4.1: GEPAS and Babelomics roadmap.

Part II

Objetives

This thesis came up with the general challenge of developing methodologies for the functional profiling of genome-scale experiments. More specifically we aimed to introduce new sources of information that increase the scope of the analyses complementing the already in-use methods in both the coverage of the annotation and the biological knowledge parcels that are explored.

This goal required the achievement of several objectives that are enumerated next:

- The introduction of new sources of annotation into the functional profiling analyses of genome-scale experiments. The sources of annotation we aimed to introduce were:
  - Biologically meaningful words associated to genes by their co-occurrence in the scientific literature.
  - Phenotype information associated to the genes by the level of expression.
  - Protein-protein interaction data.
- The development of new methodologies for functional enrichment analyses that take into consideration both the structure of the biological sequence annotation and the previous experiment design. Specifically, we wanted to generate methods that could deal with:
  - Continuous labels, that is, labels associated to the genes through a value.
  - Discrete labels with a network supra-structure.
  - Time series experiments.
  - Different experiment designs such as supervised and unsupervised problems.
- The implementation of the methods as web-based tools that could be integrated into the Babelomics and GEPAS packages for functional profiling of genome-scale experiments and analysis of microarray data respectively.
- The exploration of the possibilities of the methods developed by:
  - Performing functional profiling analysis using them.
  - Exploring the role of the protein-protein interactions into other functional classes.
  - The integration of several annotation sources to explore the variation of functional modules in cancer stages.

Part III

Materials and Methods

Chapter 5

Sources of gene/protein annotation for functional genomics

According to GOLD (`http://www.genomesonline.org`), the Genomes OnLine Database updated on the 14^{th} of July of 2008 there are 833 published completed genomes and 3887 ongoing projects. With such amount of data to deal with, biology cannot continue with the old practices for managing data based on its inclusion in books and journals. The new challenges of this new era, sometimes called post-genomics, are to build appropriate storage systems for the data and to annotate sequences following standards that facilitate an effective information retrieval system. Many attempts are now running to annotate genomes in a standardized, classified and accurate way. In the next sections we report the particularities of the annotation sources we have used during this thesis. The criteria we followed to choose them was always based on its contribution for accurate and reliable information, its coverage and its manageability.

In the introduction we already made a review of the features of the annotation sources, in here we will report how we managed this data and the technical issues we consider important to understand how the methodologies work.

5.1 Babelomics database

The project Babelomics (`http://www.babelomics.org`) has been conceived from the beginning as an integrated suite of programs with different scope but with the same internal architecture. This permits adding new modules easily and to reduce complexity improving the developers manageability of data. All the sources of annotation used are stored under a common relational database, the software used for this purpose was mysql 5.0.

A non trivial problem in molecular biology is that there is not a unique or standard name (id) for gene and protein sequences, that is, the same amino acid or nucleotide sequence is named under dozens of names depending on the institution that annotate them. The challenge in here is to know that two different ids are referred to the same sequence and cross-references are not very extended in the annotation databases. As Babelomics has as general goal of providing an easy and friendly usage to the end-user, we decided to adopt an universal index that could serve as link-node between user's reference and the annotations. We chose ensembl id for this purpose. Ensembl (Hubbard *et al.*, 2007) is a joint project between the European Bioinformatics Institute (EBI) and the Sanger Institute that came up in 1999 to develop a software system for producing, maintaining and visualising automatic annotation on selected eukaryotic genomes. Ensembl incorporates external sources of gene and protein annotation as well as the majority of the cross-references available. We make use of this facility to avoid that Babelomics users get lost under such variety of ids. Thus, users may submit any id they have and Babelomics will translated into ensembl id that is directly linked to the annotation. For the annotation sources that are not linked to ensembl ids we perform the mapping from the id provided to the ensembl gene and transcript id. This is the case of Serial Analysis of Gene Expression data, provided in HUGO Gene Name Committee (HGNC) and unigene ids, bioentities extracted form the literature, linked to HNGC ids and ppi data which is provided by several databases so several ids are used.

Although the use of an universal cross-reference has many advantages this is not free of problems. Any gene not annotated in Ensembl will be lost in the analysis. This, obviously will affect to a very small number of genes and should not affect to any general functional conclusion obtained by analysing a large and significant number of genes.

As said in the introduction, in this thesis we have used three types of annotation: discrete labels (GOs, KEGGs and BioCarta pathways), continuous labels (bioentities extracted from literature, expression levels in different tissues) and discrete labels with a supra-structure (ppis). In the next subsection we will report on the ones with special features.

5.2 Gene Ontology and Nested Inclusive Analysis

In the studies developed in this thesis where GO is the annotation source, we have applied the called Nested Inclusive Analysis (NIE) (Al-Shahrour *et al.*, 2004) in which a level in the DAG hierarchy is chosen and the genes annotated with terms that are descendant of the parent term corresponding to the level selected are annotated with such term. This increments the efficiency of the test because there are less terms to test and more genes per term.

5.3 Bioentities extracted by text-mining techniques

We have implemented two different methodologies that use bioentities associated to human genes to perform functional profiling of high-throughput experiments. These methods resulted in two web tools, Marmite and MarmiteScan, integrated within Babelomics. We will review about methods and web tools facilities in the methods subsection and in the results section respectively.

This kind of annotation was extracted using a software called AlmaKnowledgeServer which looks into the abstracts stored in Medline for gene-bioentity co-occurrences. It looks for single words but also for bi-grams (two adjacent words). This kind of terms could contain more information than both words separately, e.g. "cell cycle" gives a different kind of information than the sum of the meanings of "cell" and "cycle".

The bioentities belong to two categories, chemical compounds and disease related words. A gene is associated to a word through a score, the Z score, (Andrade and Valencia, 1998). It is calculated by the formula:

$Z_{ia} = \frac{X_{ia} - M_{ia}}{\sigma_{ia}}$ Z score for a term i in a collection of documents a

being,

N_a, the number of documents in set a

N_{doc}, the number of documents in the entire collection

X_i, the number of documents where term i appears in N_{doc}

X_{ia}, the number of documents where term i appears in N_a

$M_{ia} = N_a \times (\frac{X_i}{N_{doc}})$ the mean value for term i in collection N_a

$\sigma_{ia} = \sqrt{M_{ia} \times (1 - \frac{X_{ia}}{N_{doc}}) \times (1 - \frac{N_a}{N_{doc}})}$ the standard deviation of the distribution

The scores are based on the analysis of co-occurrences of bioentity and gene in Medline abstracts. The observed number of documents where both elements appear together and the number of documents where both appear independently are compared to an expected value based on a hypergeometric distribution. The more co-occurrences are observed in relation to the number expected the more unlikely it is that this happen by chance and the higher will be the value. Unfortunately, the absolute numbers have no meaning by themselves but can only provide an order of importance.

Table 5.1 shows a comparison between the annotation coverage of bioentities and two of the most used annotation sources (GO and KEGG). In it, we can see that although the coverage is smaller than for instance GO coverage, the number of entries, pairs of gene-label annotated, is much bigger.

The gene-label association through a score is the main difference between this kind of annotation and the more classical ones mentioned above (GOs, KEGGs

	Genes	Labels	Entries
Chemical compounds	5,832	19,605	236,466
Disease related words	5,556	6,140	204,988
Word roots	4,012	34,741	218,008
Associated genes	6,479	6,479	764,364
GO	16,423	6,193	112,634
KEGG	3,904	189	8,653

Table 5.1: Comparison of annotation coverage in human genes between bioentities (divided into chemical products, disease related words, word roots and associated genes) and two of the most used annotation sources, GO and KEGG. In columns we represent the number of genes annotated with at least one of the labels belonging to the the annotation database (Genes), the number of labels in the database (Labels) and the total number of gene-label annotations (Entries).

and BioCarta). Due to this new feature, novel methodologies for its introduction into functional genomics analysis had to be developed. As said before, the tools Marmite and MarmiteScan, part of the Babelomics suite, deal with this kind of data.

5.4 Tissue expression

In this section we will report on the datasets we have used for the annotation of high-throughput experiments results using transcriptomic data as annotation source. Their use, together with the methodology is available through the Tissues Mining Tool (TMT) module of the Babelomics suite. We wanted to introduce curated datasets that represent a broad collection of normal and cancer tissues coming from several platforms to provide a wide spectrum of exploration capabilities to the analyses. With this purpose we selected the following datasets.

5.4.1 Serial Analysis of Gene Expression (SAGE)

For our analyses we downloaded a collection of SAGE libraries from the Cancer Genome Anatomy Project (CGAP, `http://cgap.nci.nhi.gov/SAGE`) that consists of 279 human and 190 mouse high quality libraries representing a total of 29 and 26 tissues respectively and a wide range of histologies: different types of cancer, tumour associated and normal tissues. The libraries can be classified in short and long tags libraries depending if they are generated using tags of 10 or 17 bases. They already provide the tag-gene assignments.

5.4.2 Microarrays

The data we used for the integration in TMT module was generated by the Genomics Institute of the Novartis Research Foundation (GNF) and downloaded from `http://wombat.gnf.org`. Features of the dataset: Affymetrix microarray chip U133A extended with more probes.

- 79 human and 61 mouse tissues with mainly normal histology.
- Two types of normalization: MAS5 (Affymetrix method), gcRMA (Bioconductor method).

5.5 Protein-protein interactions

Finally, the last source of annotation we are going to mention in this section are the protein-protein interactions (ppis). We developed methodologies for functional profiling that use them as annotation. The methods were also implemented as a web tool integrated within Babelomics called SNOW.

5.5.1 Methodologies for ppi curation (Human interactome generation)

Our experience in compiling data to build an accurate set of human ppis showed us that the annotation in the different databases is sometimes not comparable. The approach proposed in this thesis is a modification of the one proposed by von Mering *et al.*, 2002 (see section 14.2) based in the selection of ppis detected with two different techniques.

To build a filtered interactome we took the six top categories of experimental methods described in the Molecular Interaction (MI) Ontology (Hermjakob *et al.*, 2004) plus the categories *in vivo* and *in vitro* from HPRD as reference. HPRD seems to be essential when approaching a human interactome (Mathivanan *et al.*, 2006). Every ppi in each of the datasets was annotated with these categories. Ppis verified by at least two of these methods were introduced in the filtered interactome. By using lower levels of depth in the ontology of techniques annotation we ensure that ppis extracted with experiments with similar basics that may have same biases in the detection process are not selected. 5.1 shows an schema of the application of the re-annotation process.

5.5.2 Human interactome

For generating a human interactome, we downloaded human ppi datasets from the five main public databases, HRPD (release 010107), IntAct (release 2007-04-20), BIND (release 2007-05-10), DIP (release Hsapi20070707), and MINT (release

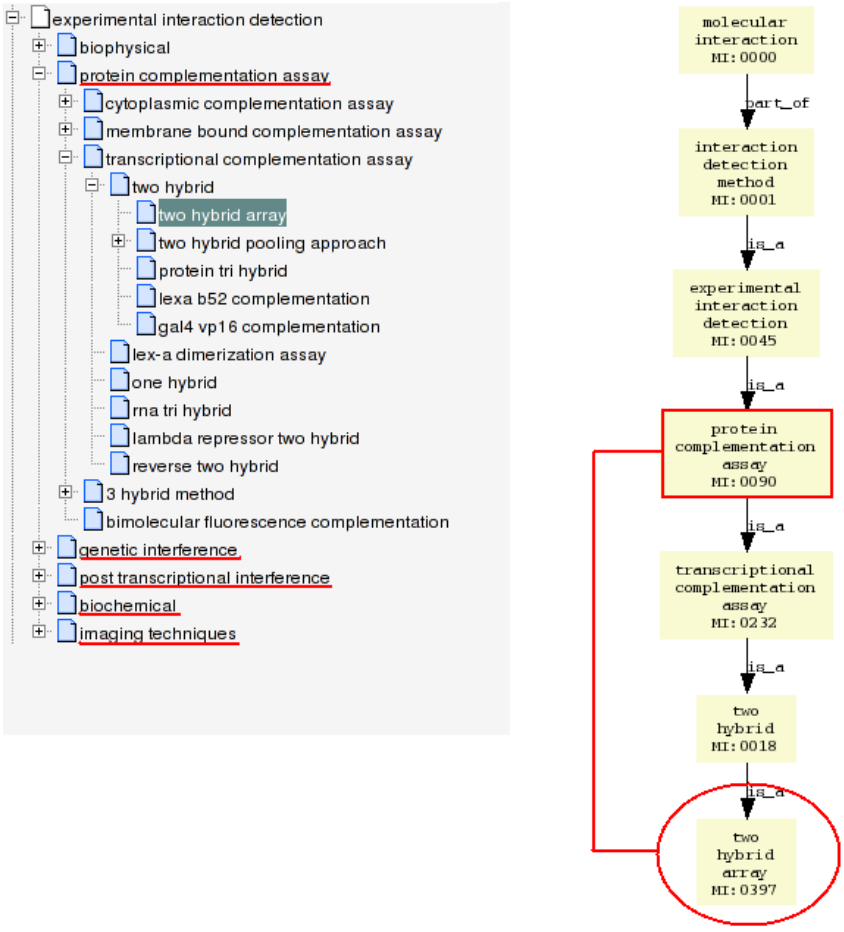


Figure 5.1: Re-annotation of ppis process schema. On the left the Molecular Interactions (MI) Ontology tree and underlined in red, the 6 top categories chosen as reference. On the right a more detailed view of the tree and a re-annotation instance, a ppi annotated with two hybrid array is re-annotated with protein complementary assay.

	Non-filtered interactome		Filtered interactome	
	Transcripts	Genes	Transcripts	Genes
Nodes	16,799	10,027	11,107	7,405
Edges	109,709	46,799	42,136	21,127

Table 5.2: Non-filtered and filtered interactomes. Number of nodes and edges in both interactomes and in genes and transcripts interactomes.

2007-04-05). Entries in databases were mapped to ensembl transcripts and ensembl genes using ensembl release 44 with the aim of avoiding duplication of same proteins with different ids.

We used this collection of ppi data to generate two different types of interactomes for both transcripts and genes: a non-filtered interactome which hold all available ppis, and a filtered interactome built using the method proposed above (section 5.5.1). Both interactomes are available in the SNOW module of the Babelomics suite, however our analyses have been performed with the filtered one only. The idea is to have an exploratory and a curated collection of ppis to be able to perform different kind of analysis. Moreover, per each type, a transcripts (proteins and transcripts have a one to one relationship) and a genes interactomes are generated because, although an artefact, typical lists of high-throughput experiments may be formed of genes. Genes have a one-to-many relationship with proteins, therefore, the topology of the network changes drastically in proteins and "genes" interactomes and mapping genes into a proteins interactome may give confused results. Table 5.2 shows the differences in number of nodes and edges between non-filtered and filtered interactomes and between genes and transcripts interactomes.

Network parameters such as connections degree, relative betweenness centrality and clustering coefficient were computed per each interactome generated and stored in a relation database (mysql 5.0).

Chapter 6

Methodologies for functional profiling

Functional Genomics and in particular functional profiling of high-throughput experiments are in a quite early stage of development. The methodologies available for this kind of analysis have been published during the last 5 years and although certainly it is one of the hot topics in computational biology, there is still a lack of novel methods that could deal with new types of annotations. Due to this fact, this thesis has a high methodological content. All the developments have been approached from a systems biology and an integrative point of view. The goal was to introduce new parcels of knowledge into the functional profiling methods as well as to develop algorithms that can be applied in a new scenario with new sources of annotation and different experiments designs.

6.1 Tissues/phenotype based profiling

The methodology presented in this section deals with gene expression measures in different tissues and histologies as annotation. Basically, it compares the distribution of the expression values of two lists of genes in different tissues and histologies and extracts the tissues in which any of the lists have significant differences.

For the analysis, we build a matrix of expression values where the genes are represented in columns and the tissues in rows. It uses an implementation of the Student test made in C programming language that finds the differences in the distribution in the two classes (lists) for every row in the matrix. It gives a p-value to every comparison that is corrected by the False Discovery Rate (FDR) method using a program written in the R programming language (see section 3.2.1.1 for an explanation on the multiple test problem) . The output of the method are the tissues in which any of the lists has a corrected p-value less than 0.05, meaning that any of the lists is significantly more expressed in the reported

conditions (tissues and histologies). The program has the option of generating the background, instead of submitting two lists, the second one (background) is generated by using the rest of the genes with expression data in the conditions selected.

The method is implemented in form of a cgi program (web program) called Tissues Mining Tool using the programming languages Perl, JavaScript and R.

6.2 Functional enrichment test using text-mining derived gene modules

This method deals with the annotation data extracted by text mining techniques explained in section 5.3. Basically, it extracts the bioentities that have a significantly higher association to a list of genes when compared to a background. The functional enrichment test carried out by this method is in many ways conceptually similar to the tests used for classical repositories such as GO, KEGG, etc. (Al-Shahrour *et al.*, 2004; Dopazo, 2006). The difference in this case is that the functional category to be tested, the bioentity, is considered to be a continuous class. Membership of a gene to a bioentity is therefore defined by a score value, which reflects the strength of the real relationship gene-bioentity. Therefore, instead of the usual Fisher's or hypergeometric (or similar) tests, we use a Kolmogorov–Smirnov test to compare the distributions of the scores of the co-occurrences between genes and bioentities for each bioentity studied to the background distribution of scores. Since all the bioentities are tested, the p-values assigned to them are adjusted by False Discovery Rate (Benjamini and Hochberg, 1995). The method returns the bioentities with a corrected p-value less than 0.05 assigned to any of the lists submitted.

The method is implemented as a cgi program called Marmite, also a module of the Babelomics suite, it is written in Perl, JavaScript and R.

6.3 Gene set enrichment analyses (Threshold-free methods)

This type of methods are more suitable for supervised analyses (analyses with previous information of class structure) such as differential expression analysis between different samples. In this type of analysis the typical two steps approach force to select a list of important genes based in a comparative parameter such as fold-change or p-value assigned to a statistic, even though the cut-off in this parameter has a statistical significance it is never directly associated to the biology, being always an artefact. Thus, as commented in the introduction, a new type of methods generally called threshold-free methods have recently appeared to avoid this problem. A representative method of this family is FatiScan (Al-Shahrour *et al.*, 2005). In this thesis we have taken FatiScan method as the model to analyse more complex experiments such a time series experiment and to be able to

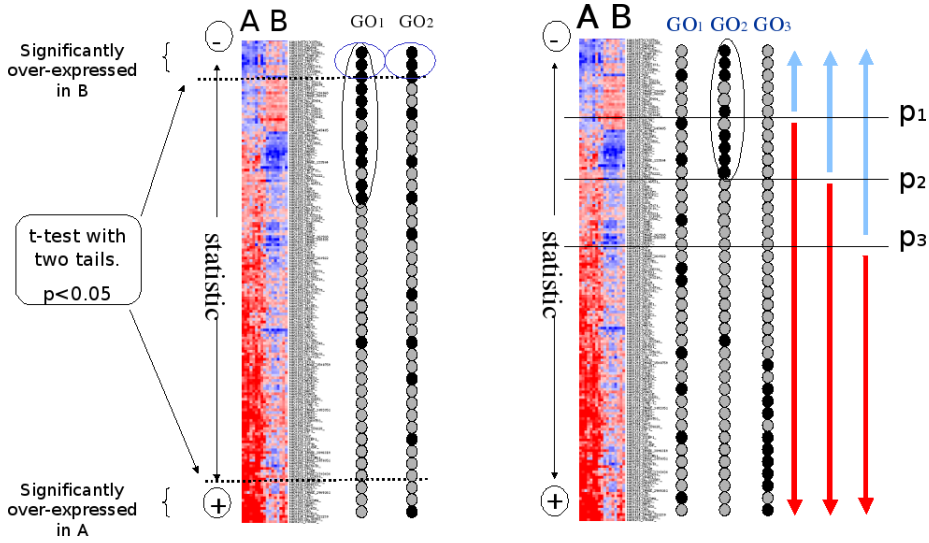


Figure 6.1: FatiScan method. On the left, a representation of the constraints of the two steps method for extracting enrichment of functional classes like GO terms. Genes in a microarray experiment with two classes (A and B) are sorted by their differential expression, from more expressed in A to more expressed in B. Two GO classes are shown. Black dots represent genes with that GO term annotated and grey dots represent genes with that GO no annotated. If we do a selection of the key genes by their p-value given by a statistical test both GOs are represented by the same number of genes in both, the list of selected genes and in the background. However the picture show that the distribution of the annotations are not the same, being GO1 genes mainly concentrated at the top and GO2 showing a quasi-homogeneous distribution. On the right we show how FatiScan can extract the GO terms with interesting distributions, the ones concentrated at the top or at the bottom, that is, the ones that have correlation with the parameter used to sort the genes, in this case the differential expression between two classes. A set of partitions is applied (p1, p2, p3 ...) and for each of them and for every GO term a two step method is performed between top and bottom genes.

use different types of annotation such as weighted associations between gene and label, see sections 6.3.2 and 6.3.3.

6.3.1 FatiScan: a segmentation test

FatiScan consists on the sequential application of the FatiGO (Al-Shahrour *et al.*, 2004) test to different partitions of an ordered list of genes. The FatiGO test uses a Fisher's exact test over a contingency table for finding significantly over or under represented biological terms when comparing the upper side to the lower side of the list, as defined by any partition. The test assigns a p-value for each functional label evaluated in every partition, the p-values are corrected due to the multiple test problem by the FDR method and finally the method extracts the functional labels with more significant p-values associated to a determined partition. See figure 6.1 for an explanation.

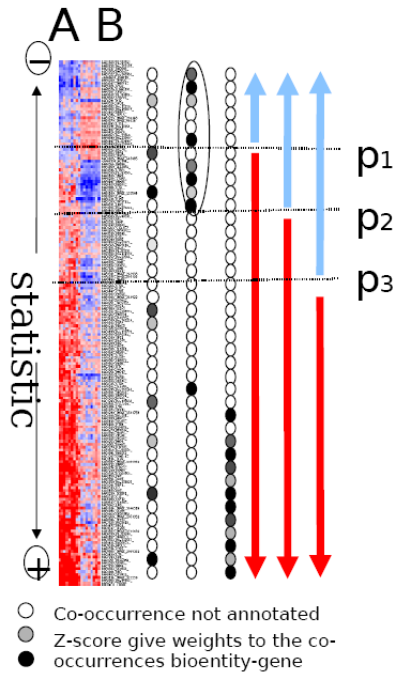


Figure 6.2: MarmiteScan method. As in figure 6.1, a differential expression analysis of a microarray experiment is shown and the genes sorted by the statistic, from more expressed in class B to more expressed in class A. The distribution of three bioentities are shown. White dots represent the genes not associated to the bioentity and coloured dots represent the genes associated to the bioentity. The scale from grey to black represent the weight of the association. For extracting the bioentities correlated to the parameter used to sort the genes, a series of partitions is applied (p1, p2, p3 ...), for each partition and each bioentity, a kolmogorov-Smirnov test is applied to test the differences in the distribution between top and bottom genes.

6.3.2 MarmiteScan

MarmiteScan is a gene set enrichment analysis to study the behaviour of blocks of genes defined by bioentities. It is carried out by means of a segmentation test similar to the one used in FatiScan (Al-Shahrour *et al.*, 2005). A pre-selection of genes is not necessary, only a ranked list is used in this test. Thus, given a list of genes arranged by any biological characteristic of the experiment (e.g. by differential expression between two types of experiments), a segmentation test is used to detect significant asymmetrical distributions of bioentities across it. Again, given the continuous nature of the bioentities, a Kolmogorov–Smirnov test is used to detect blocks of genes constitutively skewed to the extremes of the ranking and, consequently related to the biological criteria used for producing the ranking. See figure 6.2 for an explanation.

MarmiteScan is a module of the Babelomics suite and is implemented in form of a cgi program written in Perl, JavaScript and R.

6.3.3 Time Series analysis using FatiScan

6.3.3.1 Microarray Data Preparation

The data used is a microarray time series experiment of *Plasmodium falciparum* during the 48 hours of the intraerythrocytic developmental cycle (IDC) with samples taken every hour (Bozdech *et al.*, 2003). The dataset consists of 7,462 probes (70mer oligonucleotides), representing 4,488 of the 5,409 ORFs of the *Plasmodium falciparum* strain 3D7. A total of 46 time points (covering 48 hours, sampling time points in intervals of one hour; time points 23 and 29 had no data in the original dataset) were used in the study.

Microarray data used were the log-ratios of the normalized values of expression (Bozdech *et al.*, 2003). The data were arranged into a matrix of gene expression values where columns represent time points and rows represent genes. Here, time point 1 (first column) was taken as reference for the analysis. Genes with no information in this column were removed from the analysis. Then, a transformed matrix containing as many columns as time points minus one (the initial time point) was obtained by subtracting each time point from the reference time point (both are log ratios). Each column so obtained accounts for the relative differences in expression of each gene with respect to its original expression value at time 1 (or in other words, the log ratios of the gene expression values with respect to their respective value in the initial time).

6.3.3.2 Functional Analysis Using the FatiScan

The aim of the analysis is to find biological roles (according to GO annotations) that are activated or deactivated across the time series. To this end, all the columns were analyzed to detect blocks of functionally-related genes constitutively over- or under-expressed with respect to the initial condition.

The FatiScan method is applied to a list representing the differences in gene expression observed in time t with respect to the initial time, a GO term found to be significantly over-represented in top part of the list can be considered to be constitutively and significantly activated at time t. The significance of this asymmetry is obtained through a Fisher's exact test (one-tailed in this case) applied over a contingency table. The number of partitions used was of 50, which was previously shown that produces optimal results in terms of sensitivity and results recovered (Al-Shahrour *et al.*, 2005). Multiple testing effect due to the massive testing and assignation of biological terms was corrected by the widely accepted FDR. In this study FatiScan was used to search for significant GO terms from to the three main categories (biological process, molecular function and cellular component) at different levels of the GO hierarchy (levels 3, 4, 5, 6 and 7).

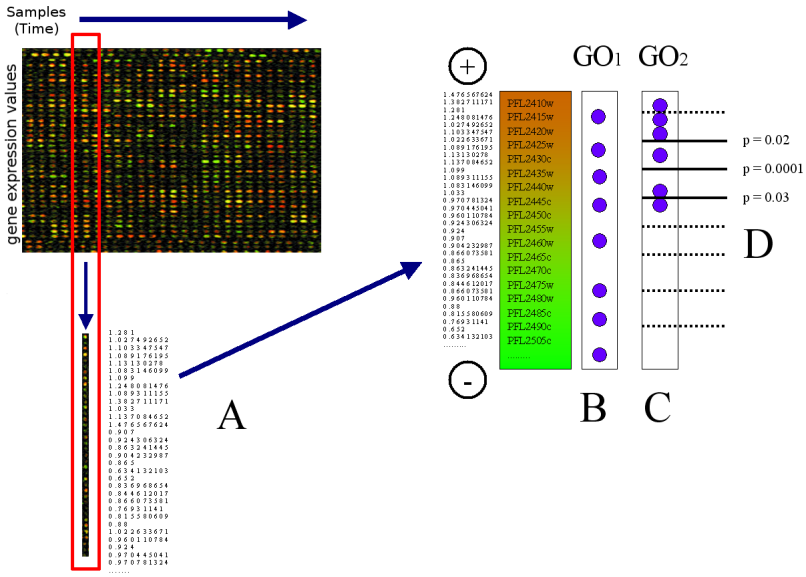


Figure 6.3: Procedure for the functional annotation of a time series. See text for an explanation.

6.3.3.3 Functional Analysis of the Time Series

Each column, corresponding to any time point beyond the initial time, was independently ordered from higher to lower values of relative expression (log ratios of the expression with respect to time 1). The positions occupied by the genes in the ranking previously obtained for each column are related to their relative contribution to the biological processes operating in this particular time point. A FatiScan analysis of each time point will render the GO terms whose corresponding genes displayed a significant coordinated high expression. These GO terms provide a detailed view on the biological roles active at the time points at which they have been detected. The dynamics of these functional roles within the cell can be analysed by plotting the GO terms significantly over-expressed along the time axis. Figure 6.3 illustrates the procedure followed. Genes at each time point are ranked from highest (top) to lowest (bottom) relative expression with respect to time 1 (Figure 6.3A). Then, for each list of ranked genes generated in any time point, the significant over-represented GO terms in the tail corresponding to the highest expression values are recorded. Figure 6.3B shows a GO term not related to high expression at this time point. Conversely, the GO term in Figure 6.3C is significantly overrepresented in high expression values. The partitions used to decide that a given term is significantly over-represented in the upper tail of the list with respect to the lower part are used for the graphical representation. The proportion of genes annotated with a significant GO term in the most significant partition is finally plotted in the graphical representation of the GO dynamics. In the example in Figure 6.3D, the most significant partition,

with p=0.001, captures the maximum divergence according to the test (with 4 out of a total of 6 terms), which would correspond to a value of 66.67%. The way in which the lists are ordered determines the hypothesis to be tested. In this case we are testing over-expressions with respect to a initial situation (point t=1), but other arrangements are possible (e.g. any point with respect to the previous one which would represent "speed" of change in functionality, etc.).

6.4 Functional enrichment using ppi data

In this section we will report on the methodologies implemented for the functional enrichment analysis using ppi data, all part of the development of a web tool called SNOW (Studying Networks in the Omic World) part of the Babelomics suite.

6.4.1 Ppis as networks

As mentioned in the introduction, ppis have the particularity of being embedded in a supra-structure. Thus, several ppis form a network. This provides the network with properties, called emergent features, that cannot be described as the sum of every ppi. The study of the topology of the network may give us clues about the role of the module in a cellular process (Yeger-Lotem *et al.*, 2004). The parameters that describe the networks can be studied using graph theory. Probably the more intuitive parameter is called connections degree and refers to the number of edges (interactions) that are connected to a node (protein). Although there are dozens of parameters that may be used, we have chosen the ones that better fit a biological meaning. The Betweenness Centrality of a node ν (CB (ν)) is a parameter that accounts for the centrality of the node ν within the graph. It is obtained from the expression:

$$CB(\nu) = \sum_{s \neq \nu \neq t \epsilon V} \frac{\sigma_{st}(\nu)}{\sigma_{st}}$$

being $\sigma_{st}(\nu)$, the number of shortest paths through a node and $\sigma_{st}(\upsilon)$, the total number of shortest paths in the graph. Relative betweenness centrality (rCB (ν)) is calculated as:

$$rC_B(\nu) = \frac{2 \times C_B(\nu)}{n^2 - (3n - 2)}$$

being n the total number of nodes in the graph.

A node with high betweenness centrality is a protein which has many shortest paths between any two other nodes passing through it. Shortest paths were calculated using the Dijkstra algorithm (Dijkstra, 1959). The action of removing that particular node from the network would cause a strong disconnection of the network. By this description, central nodes seem to be crucial in the global

compactness of the network and in the definitions of boundaries of sub-networks seen as modules of action (Girvan and Newman, 2002; Wilkinson and Huberman, 2004; Joy *et al.*, 2005). Betweenness has also been shown to be bigger in modules formed by proteins associated with cancer (Hernandez *et al.*, 2007).

Clustering coefficient of a node ν ($C(\nu)$) is a measure of connectivity that evaluates how connected is the node's neighbourhood. This parameter helps to distinguish among highly connected proteins that form star-shaped sub-networks (classical hub configuration) and proteins in a more connected area, e.g. complexes. The clustering coefficient is calculated as follows:

$$C(\nu) = \frac{2en}{n\nu(n\nu - 1)}$$

where en is the number of edges among the nodes connected to node ν, and $n\nu$ is the number of neighbours of node ν.

Other interesting features in the structure of a network are the concepts of components and bicomponents. A component is a group of nodes connected among them and a bicomponent is a group of nodes connected to another group of nodes by a single edge, which is called the articulation point (see figure 6.4). When analysing network parameters of a set of proteins with effects in the phenotype (such as gene/protein signatures of diseases or differentially expressed genes in a two conditions comparison in microarray experiments) components and bicomponents can be considered extreme examples of gene/protein modules if they can be defined as sub-networks with a higher internal connectivity than its connectivity to other modules.

We used Boost c++ graph libraries (`http://www.boost.org/libs/graph/doc/index.html`) as the software core for performing graph parameters calculation.

6.4.2 Network features evaluation

A classical experiment in Functional Genomics consists in the assessment of a functional profile to the results of a genome-scale experiment such as microarray analysis. Typically, the outcome of such experiments are lists of genes or proteins potentially relevant to the case of study because they have a common behaviour in their expression (for instance, they are over or under expressed in a disease or they have a similar pattern of expression through time when a disease is treated with a drug).

Making use of the ppi data capabilities, an interesting analysis to apply to a list of genes or proteins is to see whether it is enriched in any particular type of nodes; that is, whether it has significantly more hubs, central proteins or proteins in a very connected area compared to the complete interactome. This can be done by comparing the distribution of the connections degree, betweenness centrality and clustering coefficient respectively versus the distribution of these parameters in the set of ppis assigned as background (a curated set of the ppis

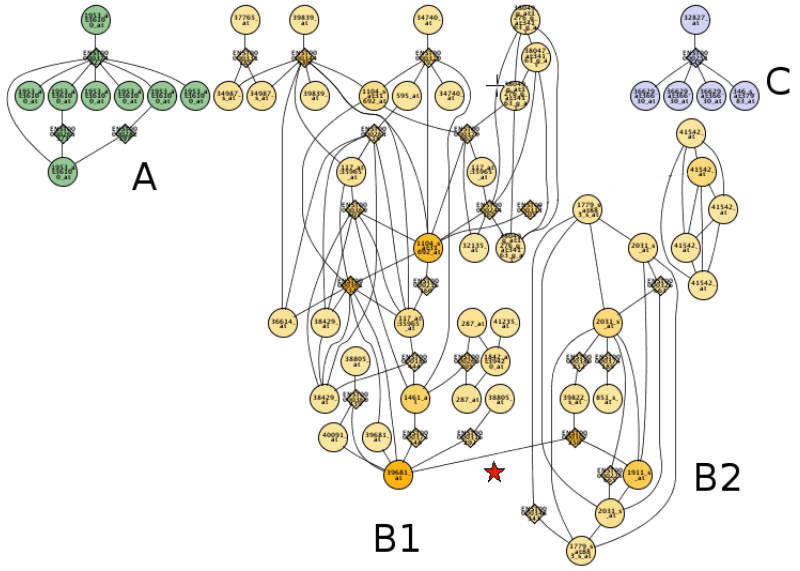


Figure 6.4: Components and bicomponents. The picture shows a network with three different components (A, B and C) coloured in green, yellow and blue respectively. Component B has in turn two bicomponents (B1 and B2) connected by only one edge highlighted with a red star.

publicly available should be representative of the complete interactome). Quite useful information about the set of proteins might be extracted by this procedure, it has been reported that cancer related proteins exhibit a higher degree of connections and centrality than the nodes in complete interactome not associated to the disease (Johsson *et al.*, 2006; Hernandez *et al.*, 2007). Nevertheless, the final aim in this kind of experiments is to seek for modules of proteins with a cooperative activity, therefore, a more holistic approach is needed. See figure 6.5 for a comparison of two very different cases when mapping proteins of a list in the interactome.

6.4.3 Methodologies to infer a sub-network

A common approach to figure out this sub-network is to calculate the called Minimal Connected Network (MCN). The MCN is the minimal network that connects a set of nodes. It is generated by the calculation of the shortest path between any two proteins in the list. When generating the MCN for functional profiling of experiments, the resulting network should be representative of the list, therefore not all the paths should be integrated in the final graph but only the ones that connect directly two of the proteins in the list plus those which connect two listed proteins through a determined number of non-listed proteins. This number of non-listed proteins that connects two of the pre-selected proteins should be small enough to keep an equilibrium between the promotion of the exploratory capabilities of this methodology and the maintenance of the accuracy

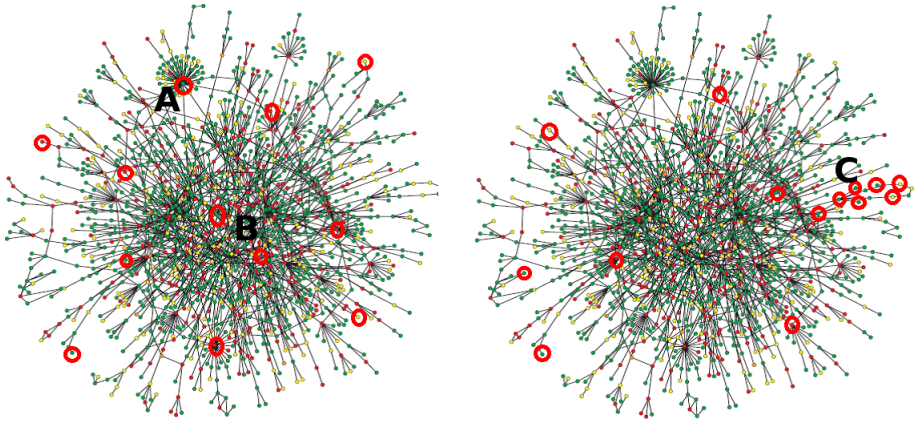


Figure 6.5: Mapping of proteins from a list in the complete interactome. Both pictures represent the interactome with proteins as dots and ppis as the edges joining the dots, the red circles highlight the nodes listed as key proteins. On the left, the mapped proteins lay into several types of nodes from external, hubs can be external as showed in A, to very central nodes (B). On the right interactome, the nodes are mainly external, which *a priori* could seem to be a less interesting situation. However, as shown in C most of the listed proteins are very close in the interactome, that indicates a possible cooperative behaviour of these proteins.

of the assignment of a network to the list.

Indeed, non pre-selected proteins by expression profiling has been reported to be related to disease due to its inclusion into a network of ppis (Xu and Li, 2006; Liu *et al.*, 2007; Chuang *et al.*, 2007). In microarray analysis it is a common practice to apply a threshold in the p-value (normally 0.05) to the selection of differentially expressed genes. This could be a constraint in the selection of important genes because many corrections have to be applied to the p-values due to the multiple testing nature of the analysis, that is, some important genes could not be reported as differentially expressed when they actually are. Moreover, there are observations that point out that important proteins in the networks such as hubs and superhubs may not be differentially expressed (Camargo & Azuaje, 2007).

6.4.4 Methodologies to evaluate a sub-network

For the evaluation of a sub-network we propose to take into account the special topological features of the biological networks. Our hypothesis is that sub-networks associated to a specific cellular activity should have a compact topology. This can be shown by having a greater distribution of connections degree and significantly less components than a sub-network integrated by random proteins that do not share functionality. We also evaluate the distribution of two other parameters, clustering coefficient and betweenness centrality, that are more related to the special features of each functional type of active networks. Thus, finding statistical significance in the different parameters points to different possible topologies of the network and the combined study of some of them may reflect

Curation	Nature	Repetitions			TOTAL
Filtered	Transcripts	X 6	X 4	X 10000	1,920,000 MCNs
	Genes	List size ranges: 3-10	Non-listed nodes allowed:	Samples	
Non-filtered	Transcripts	10-20	0		
	Genes	20-30	1		
		30-60	2		
		60-100	3		
		100-200			

Table 6.1: MCNs generated breakdown. We calculated a total of 1,920,000 MCNs divided in 2 interactomes differently curated, transcripts and genes interactomes, 6 list size ranges and 4 options in terms of non-listed nodes allowed.

the actual shape of the net. Getting significance in connections degree but not in clustering coefficient indicates a star-shaped network. The number of components and connections degree significance reflects a compact network. What is more, biological meaning of the network topology is not excluded, getting just connections degree positive results may show a set of small protein complexes and significance for betweenness centrality but not for connections degree could indicate the presence of a cascade signalling network.

This methodology use as input lists of proteins selected from a genome scale experiment for a particular reason (e.g. they co-express or they are differentially expressed under certain conditions). The aim is to find the active networks inside these lists and evaluate whether they have importance in the cooperative behaviour of the list. In other words, the methodology proposed intends to highlight the sub-networks activated in response to new stimuli and to evaluate their importance within the entity selected as the unit of study, which is the list of proteins.

We first calculate the MCN of the pre-selected proteins or genes and then we test, using the Kolmogorov-Smirnov test, each of the distributions of the MCN node parameters (connections degree, betweenness centrality and clustering coefficient) against a reference distribution.

We generated reference distributions for both filtered and non-filtered interactomes in their genes and transcripts versions. To do that we sampled ramdonly 10,000 lists of proteins and genes for each interactome and for a set of size ranges. The generation of the MCNs was done allowing 0, 1, 2 and 3 non-listed nodes. Table 6.1 summarizes the total of combinations we did to get the reference distributions.

We calculated connections degree, betweenness, clustering coefficient and number of components for every MCN generated and from each of their categories we took a random node and saved its parameters, this numbers form the reference distributions. The number of components of the MCN is compared versus a 95% confidence interval generated from the random datasets.

This novel methodology is implemented and available in the SNOW module of Babelomics suite for functional profiling of genome-wide experiments.

6.4.5 Network enrichment and networks comparison through an heuristic approach

We take the connections degree and the number of components in the MCN as indicative of a compact network. Thus, the two requisites we impose for a list to be considered as enriched in a ppi network are that the MCN that resumes it, has:

- A distribution of connections degree significantly greater (p-value < 0.05) than the connections degree distribution of the set of MCNs generated from same sized random lists.
- Less components than the 95% of the set of MCNs generated from same sized random lists.

Chapter 7

Deciphering the role of protein-protein interaction networks in the functional profiling of high-throughput experiments.

We studied the role of ppi networks as active modules in different sets of human proteins and genes either defined through transcriptome experiments or extensively used in functional profiling as definitions of functional modules.

7.1 Ppi network enrichment in Gene Ontology terms and other modules of action definitions.

A set of 8,462 lists of transcripts sharing a particular Gene Ontology term were generated using Nested Inclusive Analysis (NIA), see section 5.2 for an explanation. From those, 4,284 had less than 3 transcripts and 274 had more than 200 transcripts. Both sets were excluded from the analysis due to the difficulty of obtaining random distributions for those list sizes. The final analysis was performed with 3,904 lists (GO terms). The MCN was computed for every GO term and its node and network parameters were tested using ppi network enrichment method using the transcripts-filtered interactome. Per each GO we calculated two MCNs as the result of introducing none and one non-listed. We performed the same analysis for 146 human KEGG pathways (after excluding 41 with less than 3 transcripts and 1 with more than 200 transcripts). For a total of 313

BioCarta pathways we excluded 50 of them with less than 3 ensembl transcripts generating 263 MCNs.

7.2 Ppi network enrichment in high-throughput experiments results

A total of 665 human microarray experiment results (225 cancer, 413 no cancer, 314 up-regulated, 250 down-regulated) and 507 modules of co-expression in cancer were downloaded from L2L (Newman and Weiner, 2005). A ppi network enrichment analysis was performed for each of the sets using filtered interactome and 1 no-list node introduced.

7.3 Comparison between ppi and GO enrichment analyses

Functional enrichment analysis, using GO terms as labels, was performed applying FatiGO method (Al-Shahrour *et al.*, 2004) from Babelomics suite v2 (Al-Shahrour *et al.*, 2006) to the microarray experiments results lists and to the co-expression modules.

Chapter 8

Exploring KEGG pathways physical inter-connectivity in normal and cancer cells.

8.1 KEGGs network

To generate the KEGGs networks, both the complete KEGGs network and the phenotype specific KEGGs networks, we used as raw data the home-curated ppi interactome, details in section 5.5.1.

For the complete KEGGs network, all the ensembl transcripts within this interactome were mapped into the biochemical pathways from the KEGG pathway database. For the specific phenotype KEGGs networks we firstly generated specific ppis networks with the transcripts of each library (see next section for library details). After that we mapped the phenotype-specific ppis networks to the same KEGG pathways dataset.

In total, we generated 316 phenotype specific KEGGs networks annotated by their tissue and histology plus a complete KEGGs network. For all them we calculated their networks parameters: connections degree, betweenness centrality, clustering coefficient, number of components and bicomponents using Boost c++ graph libraries (`http://www.boost.org/libs/graph/doc/index.html`). The pairwise interactions as well as the values of the networks features were stored in xml documents.

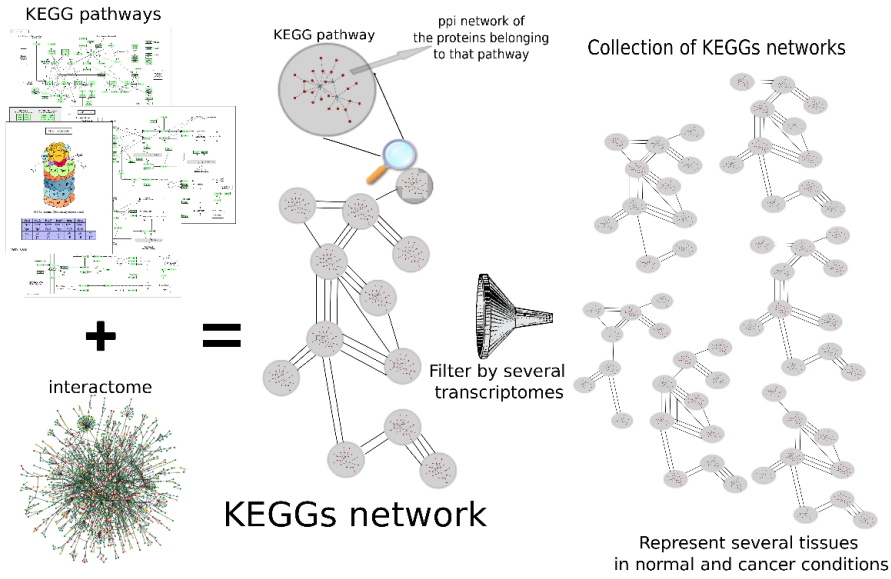


Figure 8.1: Description of how KEGGs networks collection was generated.

8.2 Transcriptomic data used to filter KEGGs networks

We used 316 human SAGE libraries (short tags) from the Cancer Genome Anatomy Project corresponding to 21 tissues with at least one normal and one cancer library. The raw data was downloaded from the CGAP web page (`http://cgap.nci.nih.gov/SAGE`) the 27th of August of 2007 and inserted in a relational database based in mysql 5.0. The correspondence between tags and genes was provided by CGAP tables, the tags are mapped to unigene and HGNC ids, we converted into ensembl transcripts using ensembl version 46. Figure 8.1 represents an schema of the generation of the KEGGs networks collection.

8.3 Connectivity Index (CI)

To measure the differences between edges weights over two sets of libraries, cancer and normal, we developed an index, the Connectivity Index (CI). Being,

ew, the weight of and edge, that is the number of ppis between proteins of a pair of nodes,

n, the number of transcripts of a library,

n(n-1)*, the maximum number of edges in a network and

N the number of libraries.

the CI of the KEGGs pair (K1-K2) is defined by the formula:

$$CI_{(K1-K2)} = \frac{\sum_{libraries} \frac{ew_{(K1-K2)normal}}{n \times (n-1)}}{N_{normal}} - \frac{\sum_{libraries} \frac{ew_{(K1-K2)cancer}}{n \times (n-1)}}{N_{cancer}}$$

To show the global trend of every tissue to gain or lose inter-connectivity in each KEGG meta-category in cancer, we summarised every CI into a categorical index, 1 for lost in cancer (CI > 0), -1 for gained in cancer (CI < 0) and 0 for no differences (CI $= 0$). After assigning one of these three values to every KEGG pair in each of the scenarios defined by the tissue and the KEGG meta-category, we extracted a global index for each of them by adding up all the categorical CIs and dividing the result by the number of KEGGs involve. Thus, we have a global index for each of the tissues and each of the meta-categories.

Part IV

Results

Chapter 9

Tools for functional profiling. The Babelomics suite

The high-throughput experiments such as DNA microarrays are snapshots of the cell stage in a particular instant. They measure global genome activity in terms of transcription. The number of probes of a typical microarray experiment is about 50,000. This size is definitely prohibitive for the traditional methodologies for data analysis in molecular biology. Therefore, their emergence has been followed by the development of databases that could deal with the management of this huge amount of data and bioinformatics tools for the extraction of relevant biological information. This thesis aims to contribute to the second aspect. We have implemented several methodologies integrated in Babelomics, a web-based suite of programs for functional profiling of genome-scale experiments.

Babelomics first release was in 2005 and since then it has been continuously evolving. New modules have been added with the aim of providing more sources of annotation and novel methodologies. Babelomics has been conceived as an integrated suite of programs directly connected to GEPAS, one of the most complete integrated packages of tools for microarray data analysis available over the web. Therefore, using GEPAS together with Babelomics, a complete analysis of a microarray experiment can be performed, from normalization to the most advanced methods for functional profiling. However, Babelomics is not restricted to microarray data. Apart from being totally platform independent, the programs are designed to accept any kind of list of genes or genes associated to a value. The criteria of the selection does not affect the methodology although it must be considered in the interpretation of the results.

A brief description on the modules and their classification available in last Babelomics release (v3.0) is given next:

- *Tools for functional enrichment analysis:*

  - **FatiGO:** performs functional enrichment analysis by comparing two lists of genes by means of a Fisher's exact test. Gene modules tested include functional criteria (GO, KEGG, BioCarta, etc.), regulatory criteria (transcription factor targets, miRNA, etc.) and chromosomal location.
  - **Marmite:** allows performing functional enrichment tests using text-mining derived gene modules. Three types of definitions are used: generic functional terms (word roots), genes associated to diseases or symptoms and genes associated to chemical compounds.

- *Tools for gene set enrichment analysis:*

  - **FatiScan:** FatiScan implements a segmentation test which checks for asymmetrical distributions of biological labels associated to genes ranked in a list. This test only needs the list of ordered genes and not the original data which generated the ranked lists. This means that can be applied to the study of the relationship of biological labels to any type of experiment whose outcome is a sorted list of genes.
  - **MarmiteScan:** extracts relevant information from a list of sorted human genes analysing precomputed gene-bioentity co-occurrences obtained by a text-mining technique. This tool finds blocks of genes more related to meaningful bioentities than what it is expected by chance.

- *Tool for tissue/phenotype-based profiling:*

  - **Tissues Mining Tool:** compares expression profiles of two lists of genes in a set of tissues. The aim of the tool is to extract different patterns of expression between two groups of genes in different tissues and histologies. In order to improve the possibilities of the analysis and to cover most of the scope of the possible experiments users are interested in, we provide data from two type of platforms, SAGE Tags and Microarray (Affymetrix) expression data.

- *Tool for functional annotation:*

  - **Blast2GO:** Blast2GO is the web counterpart of an already running Java application (Conesa *et al.*, 2005) for high-throughput functional annotation of (novel) DNA or protein sequences.

- *Other utilities:*

  - **Rosetta:** cross-reference of gene and protein ids. Rosetta includes most of gene and protein ids available.
  - **GOGraphViewer:** a viewer tool that generates joined gene ontology graphs (DAGs) to create overviews of the functional context of groups of sequences. Interactive graph visualization allows the navigation of large and unwieldy graphs often generated when trying to biologically explore large sets of sequence annotations.

In the following sections we will give a more detailed explanation of the web applications developed during this thesis, all of them integrated in Babelomics: Marmite, MarmiteScan, Tissues Mining Tool and SNOW that, although not in version 3.0, it will be available in next release.

9.1 Marmite

Marmite (Minguez *et al.*, 2007) stands for "My Accurate Resource for Mining Text" and it is placed among the modules in Babelomics that perform a two steps approach-based functional profiling of genome-scale experiments. Thus, a previous step to select the "key" genes of the analysis should be done by the user (GEPAS can be used for this purpose).

Marmite aims to characterise lists of genes by means of their co-occurrence in the scientific literature with meaningful words (bioentities) related to different biomedical aspects like drugs, chemicals or diseases. The co-occurrence between a word and a gene in a scientific paper is assumed to be indicative of some kind of association that in principle cannot be established as positive or negative. The co-occurrences among bioentities and genes are extracted from the biomedical literature by a text-mining software. Such software provides a score that evaluates the strength of the association comparing their common and solely appearances in PubMed abstracts, see Materials and Methods section 5.3 for more details.

As mentioned in the introduction, this kind of approaches are still necessary because the annotation of the genomes and proteomes is far away from being complete by means of standard terminology like GO terms, Mesh words or others. Moreover, the information that can be extracted through the bioentities cover particular aspects of the research achievements that are far out of the scope of common annotation sources like GOs or KEGGs.

However, there are also weak aspects in the extraction of meaningful association between genes and bioentities. The first one would be the lack of standard nomenclature when referring to the gene's or protein's name. Apart from the variety of ids used by the databases to annotate a sequence, the gene name may also have synonyms that have been carried out historically, but they are not recognized by any sequence database as main ids. The gene names and synonyms may also be common English words that can be misinterpreted as pointing to a gene when they are not (some examples are genes called archipelago, capicua or ebony). The writing style and word usage are also a constraint to the power of text-mining techniques because the same concept may be expressed by several words and they should be evaluated in common, not separately.

A Marmite analysis starts with the selection of a set of "important" genes. This depends obviously on the experiment design and should be done with enough statistical confidence. The list of genes selected is compared to a second list that is normally the background (rest of the genes in the genome or at least in the experiment). Users may select the type of bioentity they want for the analysis: *disease associated words* if they want to find out modules of genes acting in the

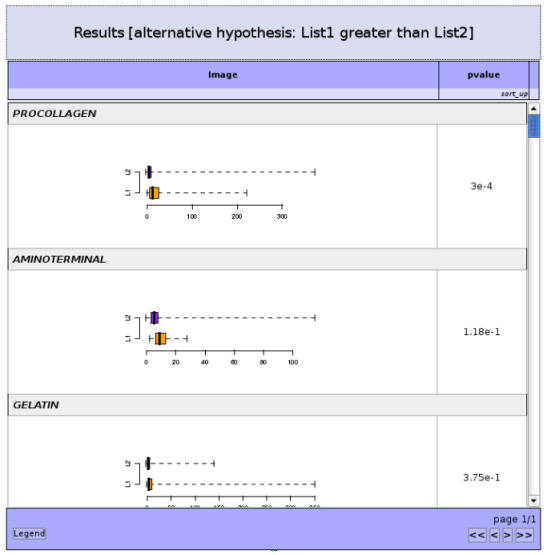


Figure 9.1: Marmite results example. Two lists of genes were submitted, these are the results for the hypothesis that score distributions of gene-bioentity co-occurrences of list 1 is greater than score distribution in list 2. Each row represent a bioentity found significantly over-represented in list 1, the boxplot with both distributions and a p-value corrected by FDR are shown.

same disease, *chemical products* for modules of genes associated for their common relation to a particular drug or compound, or *word roots*, a more speculative definition of module where the genes have association to words with the same root, e.g. catabol, apoptosi. The Marmite input web form also provides three more options to select:

- **Minimum number of genes with a score:** gives the possibility to restrict the bioentities tested to those with a certain number of annotation within the list. The speed of the method will increase and the adjustment of the p-values will be lower if we just test those bioentities with a good coverage of annotation of the list submitted.
- **Number of bioentities in the results:** controls the number of bioentities shown in the results page. The significant bioentities will always be shown but user may control to get also the ones with not significant differences.
- **Click if HUGO, HGNC, gene names:** controls how Marmite performs the mapping of the list elements. The link in the Babelomics database between genes and bioentities is made through the HUGO id (also called HGNC or gene name) because the data was provided in this way by the text-mining software. If user submits the same id as it is stored in Babelomics, the mapping is direct but if other id is provided there is a conversion step.

The results provide a list of bioentities with a significant association to any of the lists submitted, a plot of the distributions of the lists scores for each bioentity as

well as the list of the genes with their score that are associated to the bioentities in list 1 and list 2 (background). Figure 9.1 shows an example of the results of a Marmite analysis.

9.2 MarmiteScan

MarmiteScan (Minguez *et al.*, 2007) is the gene set version of Marmite. It uses the same source of annotation and usage philosophy but avoids a previous step of selection of key genes (see Materials and Methods section 6.3.2 for a detailed explanation of the methodology).

The input data for MarmiteScan is similar to the one for the FatiScan module: a list of genes with a parameter associated. The program will sort the genes using this value from greater to smaller or vice versa and extracts the bioentities that have scores distributions with a significant correlation with the ranking. Other input parameters of the application are:

- **Type of entity:** as in Marmite, users can evaluate their genes using three categories of bioentities (disease associated words, chemical products and word roots).
- **Filtering entities to test:** Parameter to select minimum number of genes with a score for an entity. Entities with less than this number in both lists will be excluded from the analysis. Default and minimum is 5.
- **Number of partitions:** Parameter to select the number of partitions that the algorithm makes to the sorted list of genes. Partitions are made based on the values associated to the genes. Users may choose values between 20 and 100.
- **Threshold p-value:** Threshold for the p-value to classify a bioentity as significant. Users can choose between 0 and 0.2 (default: 0.05).
- **Number of entities to present in the results:** Controls the number of bioentities presented in results page. Entities with significant p-values will be always shown anyway, so this restriction will never produce a lack of relevant information. Setting it to 0 means that only significant bioentities are shown.
- **Submit gene lists:** User should click this checkbox if the lists have only gene names (HGNC ids, HUGO ids, common names). The annotations are done using HUGO ids, so what MarmiteScan does is to convert any gene id to HUGO id through an ensembl id. If user provides gene names, the conversion process will be omitted, otherwise some genes might be excluded from the analysis because they match with two ensembl ids or the ensembl id match with two HUGO names.

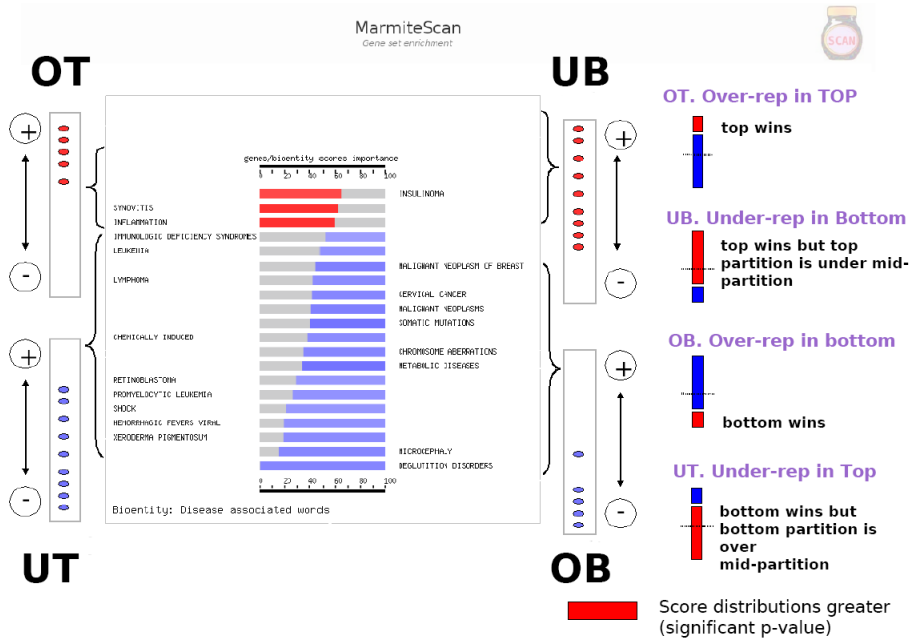
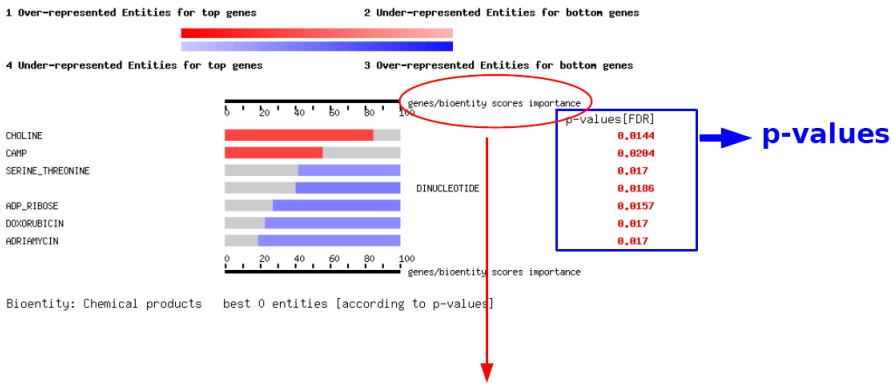


Figure 9.2: Representation of MarmiteScan results. See text for explanation.



Small formula representing the importance of the scores:

Mean of scores in Query / (Mean of Scores in Query + Mean of Scores in Reference)

Figure 9.3: More details of bioentities characterization in MarmiteScan results. See text for an explanation.

- **Do you want us to sort genes/values for you? Indicate direction:** Users may choose to sort the genes according to the value in both directions. This option may also be used to change the hypothesis to test.

The results show the bioentities with a score distribution that have any kind of correlation with the distribution of the parameter used to sort the genes. In figure 9.2 we represent the four possibilities of the correlation. The graph may be divided into four quadrants labelled as OT (Over-represented in Top), UB (Under-represented in Bottom), UT (Under-represented in Bottom) and OB (Over-represented in Top). To understand what "top" and "bottom" mean, we have to consider that MarmiteScan is a segmentation test where a set of partitions are applied to the ranked list of genes. Each partition divides the list in two and then a comparison is performed between the genes at the top part and those at the bottom part. Bioentities placed in OT quadrant are those which have a higher score distribution in top genes and the partition in which the most significant p-value has been found is over the middle partition. If the partition is under the middle partition, we cannot say that the bioentity is over-represented at the top of the list because the middle partition establishes the borderline between top and bottom but we can certainly say that the bioentity is under-represented in bottom genes (UB quadrant). The same situation is applied when bottom genes have a higher score distribution, in OB quadrant we place the ones in which the significant partition is under middle partition and in UT quadrant those with significant partition over middle partition.

Apart from the quadrant in which a bioentity is placed, there are other details that characterise each bioentity (figure 9.3):

- **P-value:** lower p-value found in the comparisons performed. There are as many comparisons as partitions are made. The p-value is adjusted by FDR.
- **Colour:** red if top genes have a higher score distribution than bottom and blue if bottom have it. The colour grade indicates in which partition the lowest p-value have been found from more to less coloured.
- **Genes/bioentity scores importance:** This indicates the importance of the scores of the set of genes in which a lower p-value have been found regarding the total values of the scores in the complete list. The formula applied is shown in figure 9.3.

9.2.1 A case of study of AML

A recent study (Stegmaier *et al.*, 2004) described a high-throughput screening methodology to test whether the action of a number of compounds in the transcriptome of cells with Acute Myeloid Leukemia (AML) reproduces the gene signature characteristic of AML differentiation to normal cells.

The gene expression data of each AML cells treated with a compound was compared to i) the expression data of the negative controls, ii) AML cells or iii)

AML cells treated with compounds that do not alter gene expression. For the comparison we applied a Student t-test to each pair of classes: AML+compound versus the corresponding control using the T-Rex tool from GEPAS (Herrero *et al.*, 2003; Montaner *et al.*, 2006, Tarraga *et al.*, 2008) to build up a list of genes ranked according their differential expression between both classes. Then, the output of this step was a set of lists of genes sorted by the t statistic and therefore by their importance in the difference between the compound action versus the AML status. MarmiteScan method was used as framework to implement a specific test to extract statistically significant chemical products related to the genes of the set of ordered genes. A Kolmogorov-Smirnov was applied to compare the distributions of the scores of the co-occurrences between genes and chemical products across the list in order to detect significant asymmetries. The p-values assigned to the bioentities were adjusted by False Discovery Rate (FDR) (Benjamini *et al.*, 1995) considering all the tests performed in the analysis.

The result of the application of the MarmiteScan algorithm to the list of genes ranked by the T-Rex program are the bioentities (chemical products) significantly associated to high gene expression in the AML+compound condition with respect to the untreated AML cells. We have found five experiments (AML+ compound) in which bioentities with a significant differential association could be detected (Table 9.1). Two out of the five compounds [*Erythro-9-(2-hydroxy-3-nonyl) adenine HCl* and *5-fluorouridine*] were found to have a gene signature similar to those that effectively differentiate AML cells to normal according to Steigmer *et al.* (2004).

In this example, the results should be understood as co-activations of blocks of genes, which have been related with chemical products through the biomedical literature, when two experimental conditions are compared (treated AML cells versus different controls). The nature of the chemical products found can give a better understanding of the biochemical processes acting in AML cells with the different treatments received. Although a detailed study of the actions of the genes annotated with the significant bioentities is out of the scope of this paper, some relevant pointers to the most interesting findings follows.

Interestingly, phosphatidylinositol has been found significant in three different experiments. It constitutes the substrate for the Phosphatidylinositol 3-kinase (PI3), a key enzyme reported of being activated in AML cells mediated by the gene Akt (Xu *et al.*, 2003) implicated in cell proliferation, cell growth and cell survival. The study of this pathway could give good clues on the processes operating in the background of these experiments. The compound fluorouridine acts by stopping DNA synthesis and thus inhibiting cell proliferation (Stegmaier *et al.*, 2004). Bioentities found for this compound may also reflect this action. It is known that glutathione has an important role in DNA synthesis, cell proliferation and apoptosis as well as in protecting cells from toxics (Wu *et al.*, 2004) such as hydrogen peroxide, which are also among the significant terms found in the fluorouridine experiment. Cyclic AMP (cAMP) promotes myeloid differentiation (Maddox *et al.*, 1988) and it has been found significant within an experiment that could not be assigned to the pro-differentiation compounds class by Stegmaier *et al.* (2004)

Compound	Bioentity (chemical product)	FDR-adjusted p-value
Erythro-9-(2-hydroxy-3-nonyl) adenine HCl	calcium	0.0349
16-ketoestradiol	calcium	0.0108
	hydrocortisone	0.0371
	cAMP	0.0121
5-fluorouridine	Sodium salicylate	0.0107
	oxygen	0.0317
	Hydrogen peroxide	0.0192
	Phorbol 12 myristate 13 acetate	0.032
	thioester	0.0328
	spingosine	0.032
	phosphatidylinositol	0.0014
	glutatione	0.0012
α-methyl-L-p-tyrosine	acetylcholine	0.0019
	noradrenaline	0.0487
	cAMP	4e-04
	dihydrotestosterone	0.034
	proteoglycan	0.0298
	calcium	0.0378
sulmazole	choline	0.0144
	cAMP	0.0204

Table 9.1: List of bioentities found to be significantly over-represented in the treatment of AML cells with an specific compound. In bold, the compounds found to have a gene signature similar to those that effectively differentiate AML cells to normal according to Steigmer *et al.*, 2004.

9.3 Tissues Mining Tool (TMT)

This module aims to give additional information about the general phenotype transcriptional profiling of the genes selected for a two steps functional enrichment analysis. Cancer and healthy tissue expression data from two different platforms (SAGE and Affymetrix) are integrated in the module to draw a general picture of the expression pattern of the list of genes as a whole (see Materials and Methods section 6.1 for more details on data and methodology).

The tool may have two different usages:

- The validation of the "gene selection" step in a functional enrichment analysis whenever the original genomic experiment has a correspondence with any of the tissues and histologies (phenotypes) provided by the tool.
- As a exploratory additional profile generator of the set of genes in different tissues and cancer states.

There are also two more ways of usage in terms of the experiment design:

- Compare two lists of genes, e.g. two lists of genes with different pattern of expression (clusters).
- Compare a list of genes versus the rest of the genes (background).

The second approach can be considered a Gene Set Enrichment (GSE)-like methodology because the list of genes is interpreted as an annotation set (functional module), just like a GO term or a KEGG pathway define a set of genes. This set of genes' expression data is mapped into the whole distribution of expression values in each phenotype (tissue + histology) to see whether it is at the heighest or lowest values of the global expression profile.

Considering all these possibilities in the analyses that can be performed with TMT, some of the typical questions that can be addressed include: the characterization of the profile of activation of the pre-selected genes in different tissues and histologies, their characterization as housekeeping genes or their association to specific types of cancer.

Some configuration possibilities can be set up to refine the analysis, they depend on the dataset chosen.

Input parameters for SAGE Tags:

- **Type of tags:** Refers to the type of tags used to infer gene expression, short (14 bases) or long (17 bases).
- **Tissues:** The set of tissues that can be selected. Each one has a set of libraries with expression values of tags that are mapped to genes.

- **Histology:** The combination of histologies that can be selected. Each library is characterized by a tissue and a histology and one phenotype may be represented by several libraries.
- **Exclude cell lines:** The libraries generated from cell lines may be excluded from the analysis.
- **Minimum number in the libraries:** This is a way to select the quality of the libraries that are going to be included in the analysis.
- **Percentage for null values accepted in libraries:** A matrix is generated with the expression values where the columns represent the genes and the rows are the libraries selected. This parameter removes from the matrix (passed to the t-test) the columns (genes) with more than the specify percentage of null values.
- **Percentage for null values accepted in genes:** Remove rows (libraries) from the matrix with more than the specify percentage of null values.

Input parameters for Microarray data:

- **Kind of normalization:** TMT provides expression values normalized by two methods: MAS5 (Affymetrix method) and gcRMA (Bioconductors method).
- **Set expression value measure for multiple probes mapping same gene:** The expression value for the genes with more than one probe mapped to them may be set up as the mean, median, greatest, lowest, percentile 25 or 75 of the expression values of the probes.
- **Tissues:** The set of tissues that can be selected.

The results provide a list of phenotypes (tissues + histology) with a p-value associated showing the differences in the expression profiles. A plot of the distributions of the lists' expression values for each phenotype as well as the list of the genes with their expression value that are associated to the phenotypes in list 1 and list 2 (background). Figure 9.4 shows an example of a TMT result.

9.4 SNOW

SNOW stands for "Studying Networks in the Omic World" and comes into Babelomics as a complement of tools as FatiGO and Marmite, introducing the power of function prediction and network structured data of ppis in the functional profiling of high-throughput experiments' results.

SNOW performs two different and complementary types of analysis to the list of proteins/genes submitted:

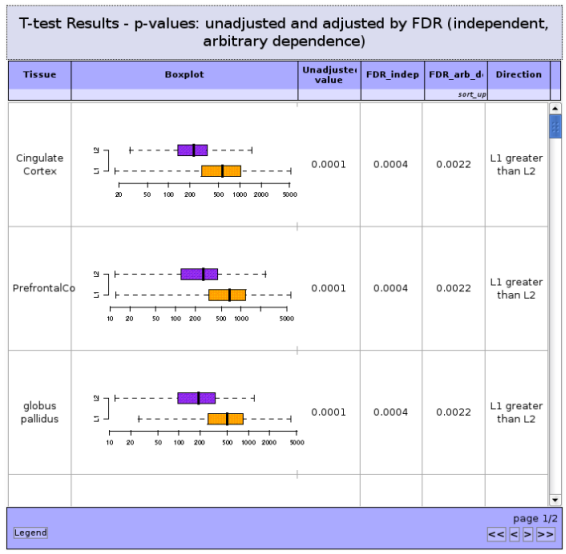


Figure 9.4: TMT results example. Two lists of genes were submitted using Affymetrix part. Each row represents a phenotype (tissue) found significantly over-represented in any of the lists. The boxplots with both distributions and a corrected p-value by several methods are shown.

- **Evaluates the role of the list within the interactome.** SNOW evaluates the global degree of connections, centrality and neighbourhood aggregation of the list by comparing the distributions of nodes connections degree, betweenness centrality and clustering coefficient respectively versus the complete distribution of these parameters into the interactome.

- **Evaluates the list's cooperative behaviour as a functional module.** SNOW calculates the MCN, the minimum network that connects the proteins/genes in the list using a user-fixed number of non-listed proteins to connect nodes in the list. The topology of this network is evaluated by comparing distributions of node, edge and graph parameters of this network against pre-calculated distributions of a set of random lists with same size range. By this, SNOW extracts information about whether the network represented in the list has more hubs, is more connected or has a more regular connections distribution than a random network. To summarise, SNOW firstly generates the functional module (MCN) and then compares it versus a background (the set of random lists). This approach is similar to other's tools for functional enrichment analysis such as FatiGO or Marmite with the difference of not having pre-annotated functional modules to evaluate, instead SNOW have to build it.

See Materials and Methods sections 6.4.2, 6.4.3 and 6.4.4 for a detailed explanation on the methods used to perform both analyses.

The input parameters for the SNOW tool are:

- **Select interactome:** The type of interactome user wants to use for the analysis, there are three options: filtered, non-filtered and own interactions, see Materials and Methods section 5.5.2 for a detailed description.

- **Submit your own interactions:** Users may perform the analysis using their own list of interactions. These should be submitted in tabulated or .sif format (compatible with other network visualization tools such as Cytoscape). If users choose this option, the evaluation of the MCN using the random lists set as background is not performed because it is a very costly computational task.

- **Proteins in interactions and list in same id?:** Only if user submits their own interactions, we ask whether SNOW needs to convert ids to be able to match your interactions with the ids in the list or they have already the same id type.

- **Maximum number of external proteins introduced:** Choose the number of non-listed proteins/genes that SNOW may introduce between two nodes to find a path. This is like choosing the maximum length of the shortest path to be able to introduce it into the MCN. The options are from 0 to 3.

- **Do you submit genes or proteins?:** SNOW will use in consequence either proteins interactome or the "genes interactome". Obviously the “genes interactome” is an artefact, see Materials and Methods section 5.5.2 for a justification.

The results of SNOW can be divided into three main parts. Two for each of the types of analysis explained before and a third one that provides a visualization facility and functional information about the MCN. Results consist of:

- **Statistical evaluation of the role of the list within the interactome:** Boxplots of the list's distributions for the genes/proteins parameters mapped into the interactome versus the whole parameters distributions in the interactome. The p-value for the Kolmogorov-Smirnov test is also provided.

- **Statistical evaluation of the MCN:** Boxplots of the list's and random's distributions for the MCNs generated from lists. The p-value for the Kolmogorov-Smirnov test is also provided.

- **Visualization and functional information about the MCN:** SNOW also provides an interactive visualization of the MCN (figure 9.5) from where users may explore the nodes and edges in the network as well as getting functional and network information of the MCN elements. A not less important part of the results is the functional information provided by means of GO terms and description of the elements of the MCN divided into components, bicomponents, listed and non-listed nodes. Moreover, the shortest

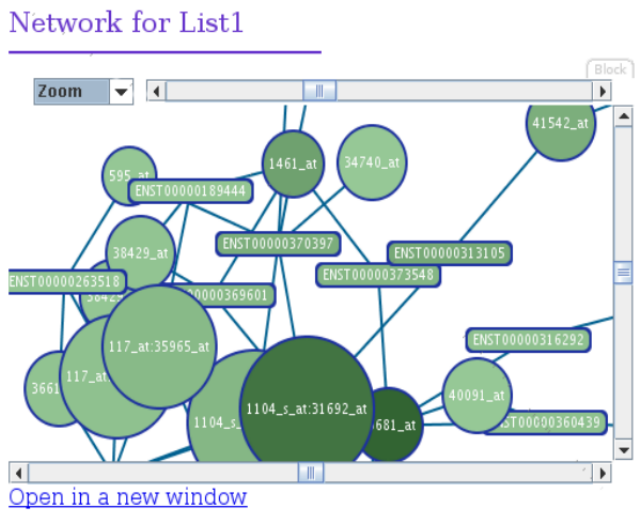


Figure 9.5: Interactive visualization of a functional module generated by SNOW. Rounded nodes are the user input proteins and the rectangles are the non-listed proteins introduced by the program. This java applet permits to zoom, move network elements, collapse edges and obtain topological and functional information about the network.

paths and the articulation points in the MCN are reported. This information will guide the user to identify the important nodes within, or even outside, the list as well as evaluate the modular functionality of the list as an entity.

A two list comparison scenario is also implemented. For it both, list nodes interactome parameters and MCN node parameters are compared using kolmogorov-Smirnov test.

Chapter 10

A function-centric approach to the biological interpretation of microarray time-series.

10.1 Time series microarrays (*Plasmodium falciparum* Intraerythrocytic developmental cycle)

DNA microarray technology has been extensively used to obtain snapshots of the expression of genes in different samples, tissues, experimental conditions, etc. While typical microarray assays are designed to study static experimental conditions, there is a class of experiments, time series, in which a temporal process is measured. Time series offer the possibility of identifying the dynamics of gene activation, which allows to infer causal relationships. Such relationships can be used to infer models of regulatory networks (Bar-Joseph *et al.*, 2003) either directly (Herrero *et al.*, 2003) or through the identification of activators and repressors (Luscombe *et al.*, 2004).

An important difference between these two types of experiments is that while static data sampled from a population (e.g. disease cases, healthy controls, etc.) are assumed to be independent, time series data are characterized by displaying a strong autocorrelation between successive points (Bar-Joseph *et al.*, 2004). Initially, time series were analysed using methods originally developed for independent data points (Friedman *et al.*, 2000; Spellman *et al.*, 1998; Zhu *et al.*, 2000). More recently, algorithms were developed to specifically address this type of data. Data analysis now address issues such as the alignment of temporal data sets (Aach *et al.*, 2001; Liu *et al.*, 2003) and the identification of differentially ex-

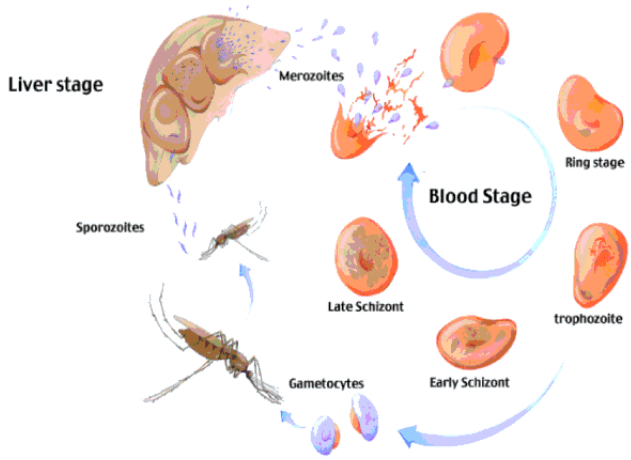


Figure 10.1: *Plasmodium falciparum* Intraerythrocytic cycle. Taken from `http://www.pasteur.fr`

pressed genes (Bar-Joseph *et al.*, 2003). Also different clustering methods specific for time series have been recently proposed. Among these, clustering based on the dynamics of the expression patterns (Ramoni *et al.*, 2002), clustering using a hidden Markov model (Schliep *et al.*, 2004) or clustering specifically devised for short time series (Ernst *et al.*, 2005).

The final aim of a typical microarray experiment is to find a molecular explanation for a given macroscopic observation, which in the case of time series is the dynamic behaviour of a system such as a cell cycle (Bozdech *et al.*, 2003; Spellman *et al.*, 1998), responses to temperature changes or stresses (Gasch *et al.*, 2000), immune response (Nau *et al.*, 2002), etc. Commonly, functional enrichment methods are used for the biological interpretation of such experiments. However, we already have seeing that these kind of methodologies it is recently being replaced in the case of supervised experiments (e.g. differential expression) by a family of methods generally called gene set methods. In this chapter, we present an application of one of this methods, FatiScan (Al-Shahrour *et al.*, 2005) to a time series obtained for the intraerythrocytic developmental cycle of the parasite *Plasmodium falciparum* (Bozdech *et al.*, 2003). The results obtained are not simple lists of genes, but the pattern of temporal activations and deactivations of the different biological roles that shape the developmental cycle of the parasite.

The asexual blood stage of the parasite *Plasmodium falciparum* (figure 10.1) causes the pathogenesis of the parasite in human, therefore understanding its gene expression profile is crucial for drug discovery and vaccine design. In blood, the parasite undergoes a 48 hours cycle characterized by three developmental forms, Ring (1-17 time points, being 1 time point equals to 1 hour), Trophozite (18-29 time points) and Schizont (30-48 time points). The mature Schizont suffers an asexual division to form up to 32 merozoites that are released to blood to invade new erythrocytes, this produces a crisis known as malaria fever. *Plasmodium*

falciparum life cycle is highly complex, involving two different hosts (mosquito, human), different tissues (liver, blood and mosquito), intra and extra cellular location and three developmental stages. Many attempts have been made to explain the gene expression profile of the intraerythrocytic developmental cycle of the parasite due to its implication in human health (Ben Mamoun *et al.*, 2001; Bozdech *et al.*, 2003).

Here we take a new approach based on the direct analysis of the dynamics of the biological roles fulfilled by the genes, as described by the GO categories. More than 40 GO terms over-represented at high expression values were obtained for some GO categories and levels, which constitute an unprecedented wealth of information on the functional behaviour at molecular level of *P. falciparum* blood cycle across time. Results for each ontology class may give different type of information. Thus, plots of GO cellular component terms show how the parasite activity moves from some cellular compartments to others (for example from nucleus in the initial stages of the cycle to membrane and host in the final stage) while molecular function and biological processes GO categories are related to biological roles of different nature coordinately played by the genes in the cell.

Here we provided only a summarized discussion of several aspects of the dynamics of the biological roles and subcellular locations at which the genes are carrying out their activities. A more detailed discussion is beyond the scope of this thesis.

10.2 Dynamics of the Biological Roles along the Cell Cycle

The plots of the different GO terms found as significantly over-represented at different times clearly illustrates the dynamics of the different roles and how the cell carries out a sequence of functional steps during its life cycle. Figure 10.2 illustrates how different biological roles switch on and off along the time points. These biological processes account for the molecular events that govern the transitions between the developmental stages of the intraerythrocytic cycle of the parasite. Just to cite a few examples, GO terms related to metabolism are found in early staged of the cycle, whereas GO terms related to signalling occur preferentially at the end, in the schizont stage.

Figure 10.3 gives information on the location, at subcellular level, where the above mentioned biological roles are taking place. It is very illustrative the fact that during the initial stages much activity occurs around locations related to replication (*RNA polymerase complex*, *nucleus*, *ribonucleoprotein complex*), while in the later stages terms related to invasion and with the interaction with the host cells are found (*host cell cytoplasm*, *host cell plasma membrane*).

In the next section a more detailed explanation of some terms in relation to the stage of the cycle in which they appear is provided. Also, in the additional information web page the three GO categories at different levels can be found. More precise terms, descendant in the GO hierarchy of the terms shown in the

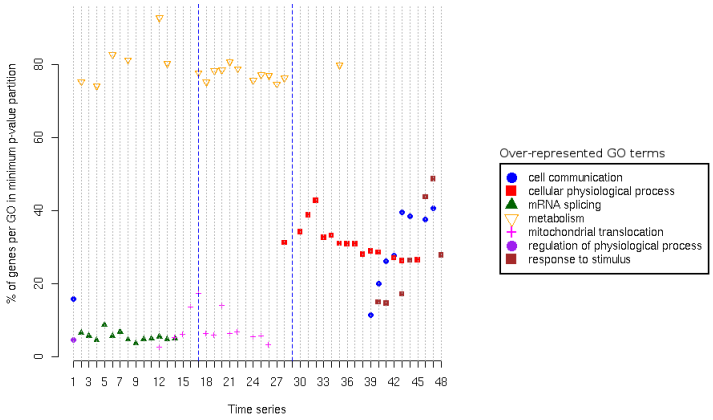


Figure 10.2: Dynamics of the biological process GO category at level 3 along the intraerythrocytic developmental cycle of the parasite. The blue vertical lines mark the transition between ring to thropozyte and from this stage to schizont, respectively.

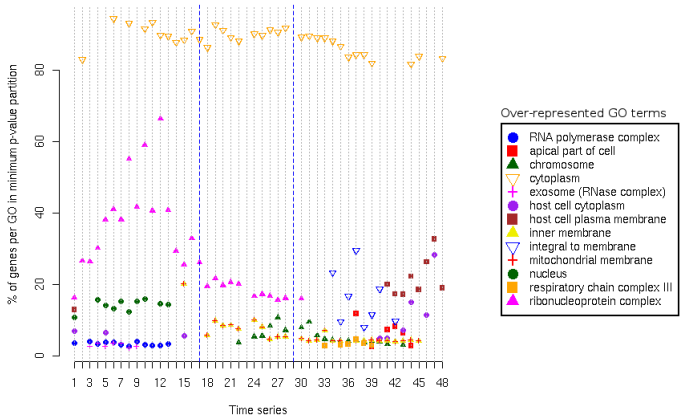


Figure 10.3: Dynamics of the subcellular location GO category at level 4 along the intraerythrocytic developmental cycle of the parasite. The blue vertical lines mark the transition between ring to thropozyte and from this stage to schizont, respectively.

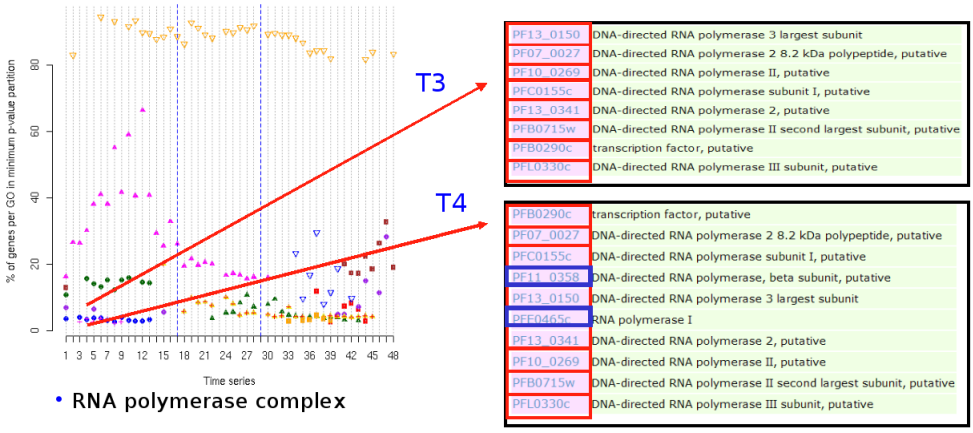


Figure 10.4: Representation of the genes activated at time points 3 and 4 annotated with the GO term RNA polymerase complex. Two new genes highlighted in blue are activated in time point 4.

figures, account with more precision for the biology of the parasite and can be found in the web page of additional information in `http://bioinfo.cipf.es/data/plasmodium`. Additionally, this page contains links to the genes in the GO categories found as significantly over-expressed. Figure 10.4 shows the type of information that can be found at `http://bioinfo.cipf.es/data/plasmodium`, for every GO category and level graph each time point is linked to the list of induced genes annotated with that GO term. Thus, if we look at two consecutive time points and the same GO term, we may see which genes are induced and repressed in such a small transition.

10.3 Biological Roles in the Different Developmental Stages

A summary of the GO terms (not all but the ones that are in consonance with previous findings as well as other relevant in this context) over-represented during the Plasmodium Intraerythrocytic Cycle follows. A complete list of the GO terms found is available in `http://bioinfo.cipf.es/data/plasmodium`.

10.3.1 Ring and Early-Trophozoite

Basic metabolism is on the top of the cell activities during this part of the cycle in which the parasite is starting the maturation process. This is reflected by a high gene activity, firstly related to transcription and then to translation, see figures 10.5 and 10.6 for details. Some GO terms found as over-represented at this

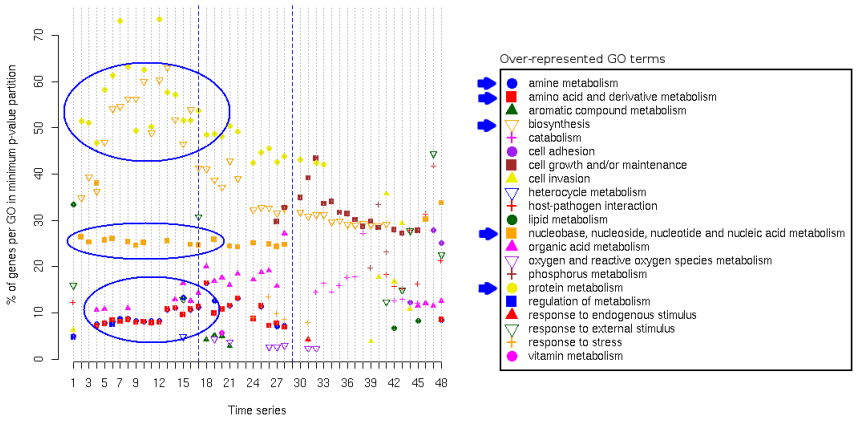


Figure 10.5: Biological process level 4. Highlighted in blue the more representative GO terms activated in Ring and early Trophozoite.

stage that support these evidences are: *transcription, DNA-dependent*, *mRNA splicing*, *RNA metabolism*, *transcription*, *RNA modification*, *RNA metabolism*, *nucleotide metabolism*, *nucleus*, *translation*, *biosynthesis*, *protein biosynthesis*, *amino acid activation*, *amino acid metabolism*, *tRNA metabolism*, *regulation of protein biosynthesis*, *regulation of translation*, *regulation of metabolism*, *organic acid metabolism*, *RNA polymerase complex* and *tRNA and amino acid metabolism*.

There is also an over-representation of terms that express the increase of the ribonucleotide biosynthesis such as: *pyrimidine base metabolism* and *pyridine nucleotide metabolism*. The presence of a high metabolic activity can be deduced from the over-representation of metabolism and other related terms like macromolecule biosynthesis and protein biosynthesis, all with a high number of genes involved. The high representation of protein biosynthesis could explain the over-representation of the *proteasome complex* term at this stage (proteosome as a quality control system). Interestingly, we found in time point 15 over-represented the terms host and host cell cytoplasm. The major stage of interaction with host comes later on in the cycle.

10.3.2 Trophozoite and Early-Schizont

We have found terms associated to DNA metabolism significantly over-represented in this developmental stage: *DNA metabolism*, *DNA replication*, *DNA replication factor*, *replisome*, *replication fork* and *chromosome*. See figures 10.7 and 10.8 for details. Although, the initial analysis of the microarray data suggested that DNA metabolism was active from the beginning of the cycle until the early schizont stage, our results seems to extend the importance of this process along almost all the schizont stage (evidences in term *DNA replication* -Biological process category- and *replisome* -Cellular component category-). Metabolism seems

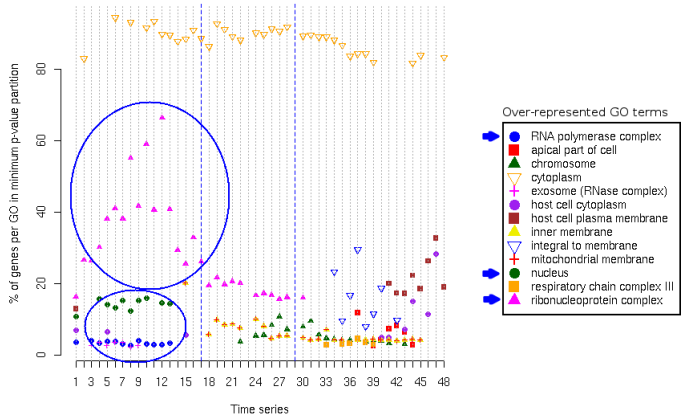


Figure 10.6: Cellular component level 4. Highlighted in blue the more representative GO terms activated in Ring and early Trophozoite.

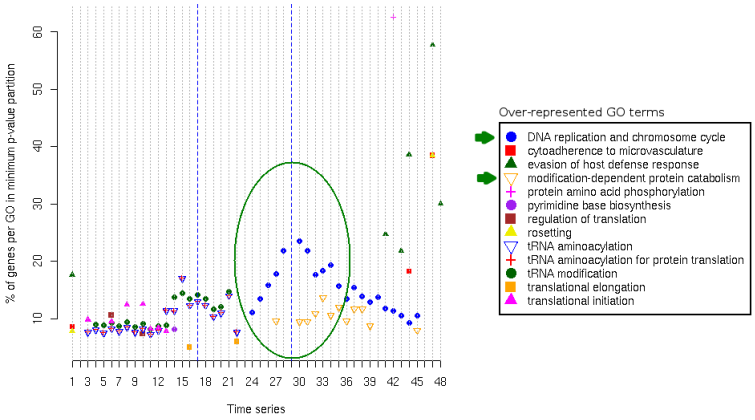


Figure 10.7: Biological process level 7. Highlighted in green the more representative GO terms in Trophozoite and early schizont.

to be also very important in this developmental stage as shown by terms such as: *carboxylic acid metabolism*, *protein biosynthesis* and *macromolecular metabolism*. Regarding subcellular location terms, cell activity has an important location in mitochondria, as shown by the terms: *mitochondrion*, *mitochondrial membrane*, *mitochondrial inner membrane*, *ion transport* and *respiratory chain complex III*. This, together with the over-representation of the terms *plastid* and *apicoplast*, supports the theory that in this period translation activity moves from nucleus to plastid and mitochondria.

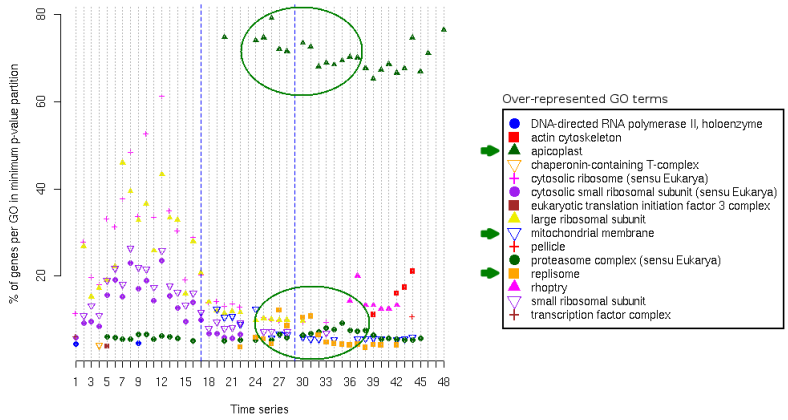


Figure 10.8: Cellular component level 6. Highlighted in green the more representative GO terms in Trophozoite and early schizont.

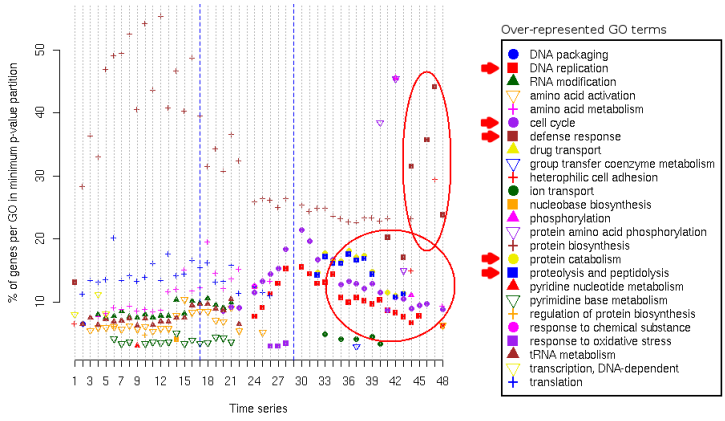


Figure 10.9: Biological process level 6. Highlighted in red the more representative GO terms in Schizont.

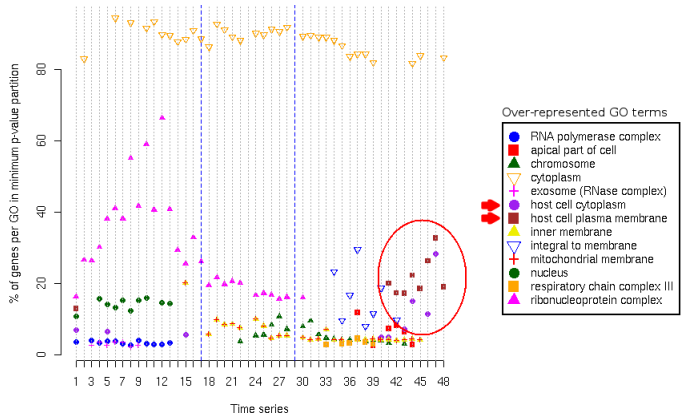


Figure 10.10: Cellular component level 4. Highlighted in red the more representative GO terms in Schizont.

10.3.3 Schizont

GO terms associated to protein catabolism become significant: *protein catabolism*, *proteolysis and peptidolysis*, *proteasome complex* and *macromolecule catabolism*. See figures 10.9 and 10.10 for more details. Other terms as *DNA replication* and *cell proliferation*, indicate that cells are in a high division activity stage. At the end of this period the main intracellular activities of the parasite are located close to membrane (as indicated by terms such as *membrane*, *host cell membrane* and *infected host cell surface knob*) and much more plasmodium specific terms associated to invasion and interaction with host become important (*cell communication*, *cell-cell adhesion*, *response to stimulus*, *response to biotic stimulus*, *response to external stimulus*, *heterophilic cell adhesion*, *defense response*, *evasion of host defense response*, *host pathogen interaction* and *cytoadherence to microvasculature*). More specific terms are in consonance to previous analysis indicating that proteases, kinases and actin-myosin motors play an important role in invasion (*kinase activity*, *myosin*, *actin cytoskeleton* and *peptidase activity*) The activity also increases in the plastid genome.

10.3.4 Early Ring

Invasion-specific biological activities still remain significant in the earliest hours of Ring stage as shown by GO terms such as: evasion of host cell response and host cell plasma membrane. In the original paper (Bozdech *et al.*, 2003) the authors alert on the possibility of some contamination from the previous stages affecting to the ring stage, so the results obtained for this stage must be taken carefully. Terms such as *cell communication*, *cell invasion*, *response to stimulus*, *cell adhesion* and other similar become visible already in the very first time points of the analysis.

Chapter 11

Deciphering the role of protein-protein interaction networks in the functional profiling of high-throughput experiments

This thesis focus on the search of functional modules in a transcriptomic context. We have developed methodologies available through bioinformatics tools and performed analysis using them to find functional classes that are induced or repressed as a block under certain conditions as deduced from transcriptomic analyses. Classically, the functional modules are sets of proteins or genes predefined by their belonging to common annotation classes like GO terms or KEGG pathways. One of the challenges for introducing ppi data into functional profiling of high-throughput experiments is that such functional modules are not defined as discrete entities but as part of a global network of pairwise interactions (interactome).

To extract the active modules from the interactome we have to consider each particular cell stage mainly because, contrarily to the nature of the genome, the proteome is not a static entity. It depends firstly on the transcriptome but also on the post-transcriptional and post-translational regulatory events. Although the regulation at transcript and proteins level is very important (post-transcriptional and post-translational regulation) the transcriptome in combination with the interactome has been used as an aproximation of the proteome to infer gene function (Ideker *et al.*, 2001), to extract signatures to predict disease phenotypes (Camargo & Azuaje, 2007; Lee *et al.*, 2007; Liu *et al.*, 2007; Chuang *et al.*, 2007) as well as to detect possible drug targets by inferring topological features of particular classes of genes (Wachi *et al.*, 2005; Johsson and Bates, 2006).

The SNOW tool (section 9.4) and its methodologies concentrate on searching ppi networks as functional modules in transcriptomic experiments. As a complement of that, we wanted to explore the role of ppi data firstly within more classical functional classes like GO terms, KEGG pathways or BioCarta pathways as well as in co-expression modules in cancer studies and secondly within lists of induced and repressed genes taken from microarray experiments.

11.1 Ppi networks in Gene Ontology terms

GO terms are typically used as a standard set of pre-defined classes to perform functional enrichment analysis. Here we wanted to check how ppi networks contribute to the formation of functional classes like GO terms. With this aim, we applied the methodology implemented in SNOW for testing the cooperative behaviour of a list of genes or proteins in ppi terms (see Materials and Methods section 6.4.4 for details) to the collection of lists of genes sharing a GO term as described in Material and Methods section 7.1. We focused on human data.

The analysis was performed using our home-made filtered interactome in two conditions: allowing none and one non-listed transcripts. So two MCNs were generated and evaluated for each GO term. The distribution of the connections degree (degree), betweenness centrality (betweenness) and clustering coefficient of the nodes in each of the MCNs generated was compared using a Kolmogorov-Smirnov test versus the distribution of the same parameter in a set of 10,000 MCNs generated from a same size range set of lists populated with random proteins/genes. For each of the parameters comparisons, two possible situations were reported:

- **Positive results.** The distribution of the parameter in the MCN found within the GO term is significantly higher than the distribution of the parameter in the random set of lists of genes with a p-value less than 0.05. For the evaluation of the components we considered a positive result when the number of components of the MCN is below than the lower value of the 95% confidence interval calculated from the MCNs extracted from the random lists.

- **Negative results.** The distribution of the parameter in the MCN found within the GO term is significantly lower than the distribution of the parameter in the random set of lists of genes with a p-value less than 0.05. For the evaluation of the components we considered a negative result when the number of components of the MCN is greater than the higher value of the 95% confidence interval calculated from the MCNs extracted from the random lists.

The percentage of the positive and negative results was calculated for each network parameter, complete results are shown in table 11.1. Two main conclusion can be extracted at first sight:

		% of lists with p-value < 0.05	
network parameter(s)	direction of comparison with random	0 non-listed genes allowed	1 non-listed gene allowed
betweenness	greater	10.12	36.53
	less	0	0.26
degree (connections)	greater	36.92	71.52
	less	0	0
clustering coefficient	greater	16.01	22.21
	less	0.03	0
bet + degree + cl. coef.	greater	6.1	15.11
	less	0	0
components	greater	0.03	0
	less	25.44	51.92
components (less) + degree (greater)		21.11	47.72

Table 11.1: Percentage of lists of genes that sharing a Gene Ontology term that have a positive result in the comparison of their parameters distributions versus random lists distributions.

- All the network parameters' comparisons give a higher percentage of positive results, that is, GO terms have better network features than random lists of proteins while the contrary rarely happens. All the percentages of negative results are zero or nearly zero. From this, we could say that ppis play a special role within functional classes defined by GO terms.

- The introduction of a non-listed node in the MCN results a better performance in the evaluation of all the network parameters of MCNs. Basically the results show that approximately the positive results increase to the double when allowing a single non-listed node.

In a more detailed observation of the results we can see that there is a different performance of the network parameters evaluated, being connections degree and number of components the two parameters with higher distribution differences followed by clustering coefficient and betweenness (when a non-listed node is allowed, betweenness show a greater distribution difference than clustering coefficient). Both, connections degree and number of components indicate big-sized networks (in relation to number of genes in the set) while betweenness and clustering coefficient are more related to the topology of the network. Therefore, although betweenness and clustering coefficient have a lower percentage of positive results, this does not reflect a lack of an active network but a network with a special topology, e.g. a signalling cascade network is a network with a low clustering coefficient distribution because its nodes are not in very connected areas.

The jointly observation of the three node related features (betweenness, connections degree and clustering coefficient) results in a decrease of the percentage of positive results. The percentage of GO terms with those three parameters distributions significantly higher than the distribution taken from the random lists of proteins is almost half than the lower percentage for a single parameter, betweenness, showing that there is not much overlapping among some of these three comparisons. Nevertheless, the percentage of positive results of number of components and connections degree almost remain the same as the lower percentage presented by the number of components, that gives the lower percentage of these two parameters, showing a great overlapping between them.

11.2 Ppi networks in other functional classes and differentially expressed lists of genes

To extend the conclusions of the study of the role of ppi networks in GO terms to other pre-defined functional classes we performed a massive analysis of lists of human genes and proteins taken from microarray experiments, co-expression modules in cancer, GO terms, KEGG pathways and BioCarta classes. The aim was studying how ppi networks are spread in different types of lists and in some classical sources of annotation normally used for functional profiling. Table 11.2 shows the percentages of the lists in each category that presented a positive result in each of the analysis performed. As explained in the Materials and Methods section 7.1 the procedure was to get every list and calculate its MCN allowing the inclusion of a non-listed node into the network. We used our home-curated human interactome. The distribution of the connections degree (degree), betweenness centrality (betweenness) and clustering coefficient of the nodes in each of the MCNs generated was evaluated in same terms as we did with GO terms, explained in previous section.

The lists of genes taken from microarrays experiments were lists of genes over and under expressed in a wide variety of phenotypes. We subdivided them into cancer, no cancer and up-regulated, down-regulated to test whether there was an association of any of the network parameters evaluated with some of these subclasses. The modules of co-expression in cancer were also taken from microarrays experiments, see Materials and Methods section 7.2 for details.

Figure 11.1 shows the complete results for the analysis. The first observation that can be done to the results is that KEGG pathways have the highest percentages of positive results in all the comparisons except in connections degree where GO terms present a much higher rate of positive results than any of the other classes. GO terms is the second functional class in the ranking of positive results percentages followed by BioCarta pathways and modules of co-expression in cancer, these two with very similar levels. Surprisingly, KEGG pathways and BioCarta, although performing both quite well, do not have similar percentages as expected by their definition (both represent biological pathways). Always BioCarta have lower rates. Lists of differentially expressed genes in microarray

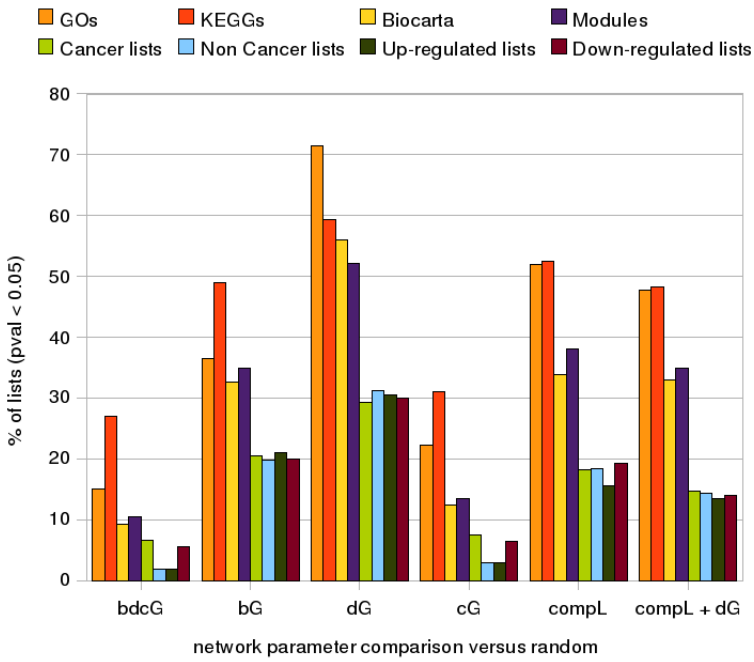


Figure 11.1: Percentages of lists in every category with a significant p-value compared to random distributions. The comparisons performed are: bdcG (betweenness, connections degree and clustering coefficient greater than random), bG (betweenness greater than random), dG (connections degree greater than random), compL (number of components less than random), compL+dG (number of components less than random plus connections degree than random).

	Gene Ontology	Modules	Cancer Lists	Non-Cancer Lists	Up-regulated Lists	Down-Regulated Lists	Biocarta	KEGGs
bdcG	15.11	10.5	6.67	1.9	1.91	5.6	9.27	26.9
bG	36.53	34.9	20.44	19.85	21.02	20	32.59	48.97
dG	71.52	52.1	29.33	31.23	30.57	30	55.91	59.31
cG	22.21	13.4	7.56	2.91	2.87	6.4	12.46	31.03
compL	51.92	38	18.22	18.4	15.61	19.2	33.87	52.41
compL + dG	47.72	34.9	14.67	14.29	13.38	14	32.91	48.28

Table 11.2: Network parameters evaluation in different types of sets of genes. Lists of genes taken from differential expression analysis of microarray experiments (cancer lists, non-cancer lists, up-regulated lists and down-regulated lists), modules of co-expression in cancer also taken from microarray experiments and genes belonging to the same annotation class (GO terms, KEGG pathways and Biocarta pathways). In rows we show the network topological parameters evaluated, bdcG (betweenness, degree and clustering coefficient have a significantly higher value than random sets), bG (betweenness significantly higher than random), dG (degree significantly higher than random), cG (clustering coefficient significantly higher than random), compL (number of components below the 95% confidence interval of random sets) and compL + dG (number of components below the 95% confidence interval of random sets plus degree significantly higher than random) . The value in the cells is the percentage of the set of lists with a p-value less than 0.05 compared to networks generated using same size random lists.

experiments are at the bottom of the ranking, showing much lower rates of positives results in all the comparisons. The classification of differentially expressed lists into four categories does not show differences among them but in the clustering coefficient comparison where cancer and down-regulated lists show a higher percentage indicating that their networks must have a higher interconnectivity.

Comparing the performance of the different network features evaluated, connections degree is the parameter that showed more positive results, as observed previously, followed by number of components, betweenness and clustering coefficient. The same conclusions as in GO terms analysis for the results of combination of evaluated parameters can be made.

11.3 Comparison between ppi and GO enrichment analyses

From the previous studies we may conclude that ppi networks do have an important role in the conformation of functionally related blocks of genes. Even in genes detected to be differentially expressed in a transcriptomic experiment, that may not be involved in a single activity but in more than one, we could detect modules of action using ppi data.

After this, we wanted to compare the performance of this novel methodology with a more classical approach for functional profiling of high-throughput

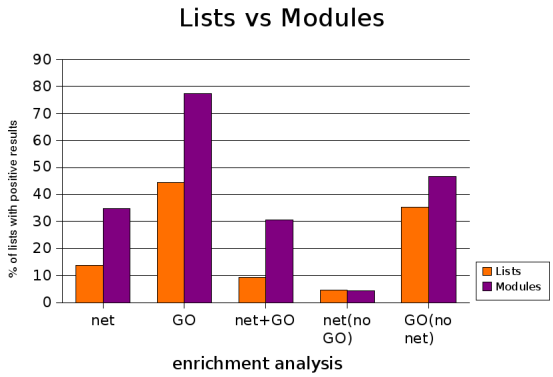


Figure 11.2: Comparison of enrichment analysis methods performance. Percentages of lists with: net (SNOW positive results), GO (FatiGO positive results), net+GO (SNOW + FatiGO positive results), net (no GO) (SNOW positive results and FatiGO negative results), GO (no net) (SNOW negative results and FatiGO positive results).

experiments. The method selected was FatiGO (Al-Shahrour *et al.*, 2004) as one of the leading representatives of Gene Ontology enrichment analysis. The input data was the set of differentially expressed genes resulting from microarray experiments and the co-expression modules in cancer used in previous section, see details in Materials and Methods section 7.3.

To be able to compare them we had to define what was a positive results for each methodology. In the case of FatiGO, a positive result was taken as the retrieval of at least one GO term significantly over-represented. For SNOW the challenge was to decide when there is an enrichment in terms of a ppi network. As explained in the Materials and Methods, section 6.4.5, we propose a heuristic criteria consisting on taking the connections degree and the number of components in the MCN as indicative of a compact network. Therefore, to be able to determine whether we have a ppi network enrichment we require positives results (as defined in section 11.1) in both connections degree and number of components evaluations.

The results of the comparison of both analyses are shown in figure 11.2. As pointed before, modules of co-expression present more positive results using both FatiGO and SNOW methods. Generally, FatiGO finds more functional modules than SNOW indicating that GO terms still define more functional classes in transcriptomic experiments. We found that nearly 5% of the performed analyses gave SNOW positive results and FatiGO negative results, that is, the functional modules acting in those situations are defined by ppi networks but not by GO terms. Interestingly, in this case both (differentially expressed lists of genes and modules of co-expression in cancer) presented same percentages.

Chapter 12

Exploring KEGG pathways physical inter-connectivity in normal and cancer cells

This study continues with the goal of learning how ppis are spread through functional classes, contributing to their definition as modules of action. The aim is to integrate several sources of annotation in order to refine the biological information extracted from transcriptomic analyses. With this purpose we designed an experiment to explore how the relationships between biological pathways, as defined by the KEGG pathways database, evolve from normal tissues to their cancer stages.

In the Introduction we already introduced KEGG pathways as one of the leading collections of biochemical pathways available. In this study we mapped the ppi data from our curated human interactome (see Materials and Methods section 5.5.1 for details on its generation) into the human KEGG pathways. The result is a network of KEGGs where the nodes are the pathways and the edges represent the interactions events between proteins belonging to the nodes (KEGGs). An edge between two KEGG pathways (e.g. KEGG1 and KEGG2) represents at least one physical interaction between a protein belonging to KEGG1 and a protein belonging to KEGG2. Thus, this two KEGGs have as many connections (weight of the edge) as interactions events occur between any protein within KEGG1 and any protein within KEGG2.

The entire collection of pairwise edges between KEGGs is called KEGGs network. This novel entity is a valuable representation of the cell functionality in terms of the integration of two very different types of information: the biochemical pathways acting in the cell and the physical interactions between their constituent proteins. In this study we have addressed both, the description of this general picture and the exploration of the changes it suffers in several scenarios represented by different phenotypes.

12.1 KEGGs network description

A KEGGs network can be represented as an undirected graph. The features we use to describe such a network are:

- **Number of nodes.** Number of KEGG pathways represented in the network.
- **Number of proteins in a node.** The number of proteins annotated with a particular KEGG pathways (node).
- **Node internal ppi network.** The set of proteins annotated with a KEGG form also a network (undirected graph) where the nodes are the proteins and the edges are the physical interaction events between those proteins.
- **Auto-interactions degree of a node.** The number of interactions between the node and itself, that is, the number of edges in the node internal ppi network or the number of ppis that occur within the set of proteins annotated with a particular KEGG.
- **Connections degree.** Number of edges of a node. An edge is defined as one or more ppi occurring between the proteins of two nodes (KEGGs). Even if there are more than one ppi between the two proteins sets, the edge is summarised as a single event, this extra information is given in the weight of an edge.
- **Weight of an edge.** An edge is defined by two nodes (KEGG1 and KEGG2), its weight is the number of ppis that occur between any protein that belongs to that KEGG1 and any other protein that belong to KEGG2.
- **General graph features.** The KEGGs networks may also be described as an undirected graph using parameters referred to their nodes such as betweenness centrality or clustering coefficient and parameters referred to the whole graph structure such as number of components or number of bicomponents. See Materials and Methods section 6.4.1 for definitions.

12.1.1 Characterization of KEGGs network

By this analysis we are re-organizing the original ppi network into a reduced network with less nodes and less edges and with new features such as the edges' weights. In other words, we are introducing a new level of abstraction or dimension: the KEGGs as nodes of a network of physical interactions. Thus, we still have the same data as in the interactome but with a new structure. The first question we addressed was whether this new structure maintained the same global features of the original ppi network.

The degree distribution $P(k)$ gives the probability that a particular node has k edges. The scale-free networks, also called small world networks, are nets where a few nodes act as "highly connected hubs" (high connection degree) and

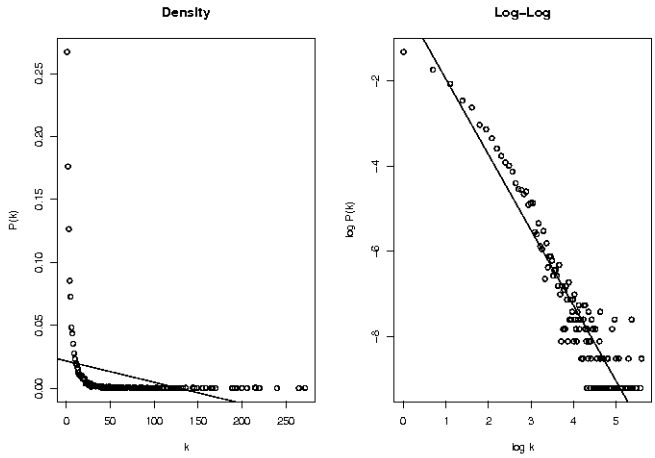


Figure 12.1: ppis interactome P(k) distribution

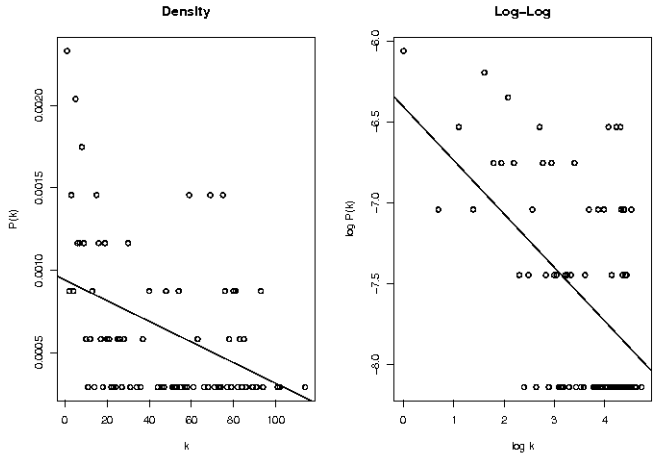


Figure 12.2: KEGGs network P(k) distribution

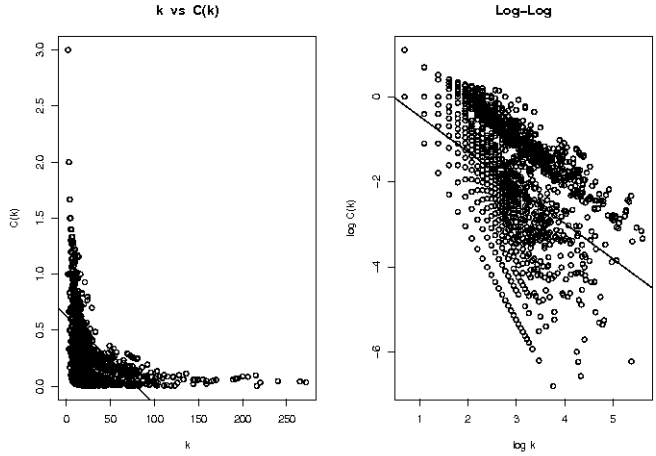


Figure 12.3: ppi interactome C(k) distribution

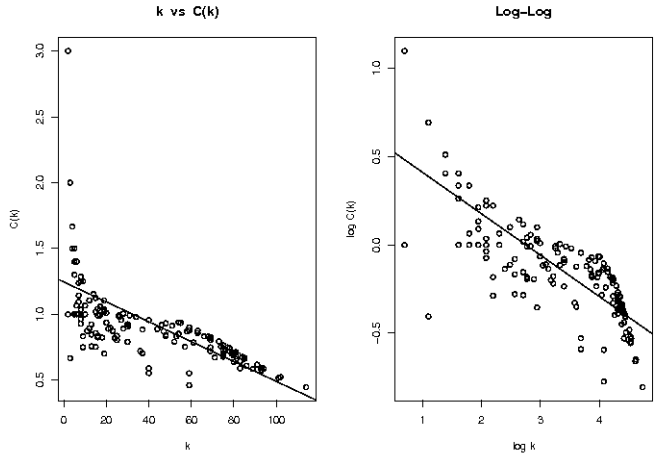


Figure 12.4: KEGGs network C(k) distribution

the rest of nodes are of low degree. Their most important characteristic is that their degree distribution follows a power law distribution defined by $P(k) \sim k^{-\gamma}$ where the probability ($P(k)$) that a node in the network connects with k other nodes is roughly proportional to $k^{-\gamma}$ ($\sim$ means proportional). The value of γ determines many properties of the network. The smaller this value, the more important is the role of the hubs. For $\gamma > 3$, hubs have not relevance. In general, the properties of the scale-free networks are valid for $\gamma < 3$ (Barabasi and Oltvai, 2004).

The values calculated for γ for our curated interactome and for the KEGGs network were 1.532 and 1.166 respectively, meaning that both interactomes may be defined as scale-free networks. This was already reported for ppi networks (Barabasi & Albert, 1999) but it was important to corroborate that KEGGs networks behave also as the majority of the biological networks studied (Barabasi & Albert, 1999). Figure 12.1 and 12.2 show the graphical representation of this distribution in both networks.

Another feature used to describe networks is the average clustering coefficient <C> that characterises the tendency of the nodes in a network to form clusters. As <k>, average of number of edges of a node, <C> depends on the size of the network, that is, on the number of nodes and edges of the network. C(k) is the function that defines the average clustering coefficient of all nodes with k edges. C(k) as P(k) are size independent so they are used to do the network classifications. Figures 12.3 and 12.4 show the graphical representation of function C(k) for interactome and KEGGs network, both having similar regression curves.

12.1.2 Role of KEGGs within the network

Once we have categorized the KEGGs network as a scale-free network with the implications that this have in its topology properties, we wanted to extract the nodes (KEGGs) with special features within the network. To assess this, we calculated connections degree, betweenness centrality, clustering coefficient and number of auto-interactions for each node in the network. Apart from the rankings of KEGG pathways according to their network parameters that shows which ones have more connections or are more central, etc. (data not shown) we considered that a more interesting analysis would be to use the annotation of KEGG pathways into meta-classes as more general classes to summarise the net in terms of its biological activity. The meta-classes are the general categories in which the KEGG database classify the pathways (e.g. the KEGG pathway *Inositol metabolism* belongs to the meta-classes *Carbohydrate Metabolism* and *Metabolism*).

We applied FatiScan method (see Materials and Methods section 6.3.1 for explanation) to seek into the lists of KEGG pathways, ranked according to each of the networks parameters calculated, for the functional meta-classes that are not randomly distributed along the list, that is, their distribution is correlated to the parameter distribution.

Figure 12.5 shows the FatiScan result for the lists of KEGGs pathways sorted

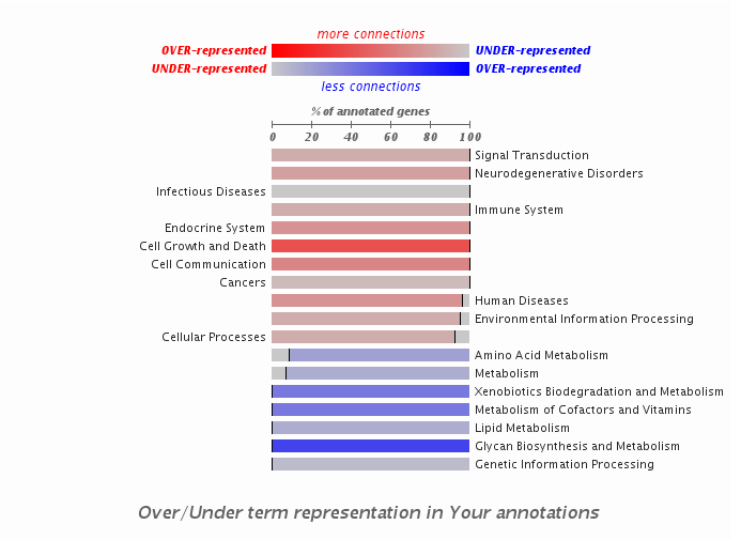


Figure 12.5: FatiScan results, connections degree.

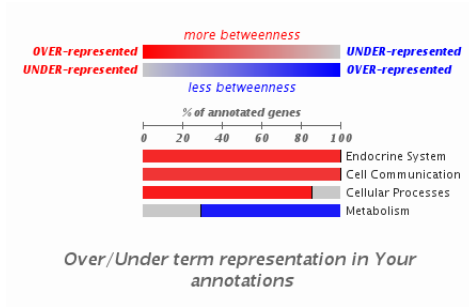


Figure 12.6: FatiScan results, betweenness centrality.

by their number of connections, from more connected (top) to less connected (bottom). The meta-classes significantly over-represented at the top of the list are *Infectious Diseases, Endocrine System*, *Cell Growth and Death*, *Cell Communication, Cancers* and *Cellular Processes*. They summarise the hubs functionality in the network together with a second set of meta-classes that, although not associated to high values of connections, they are under-represented in the lower part of the list (with less connections) that are: *Signal Transduction*, *Neurodegerative Disorders*, *Immune System*, *Human Diseases* and *Environmental Information Processing*. We can see clearly two main groups in the more connected nodes, those associated to signalling (*Cell Growth and Death*, *Cell Communication, Signal Transduction*, *Environmental Information Processing*) and those associated to diseases (*Neurodegerative Disorders*, *Immune System*, *Human Diseases*) in which many of the proteins involved in signalling are participating.

On the other hand, the classes associated to low values of connections are

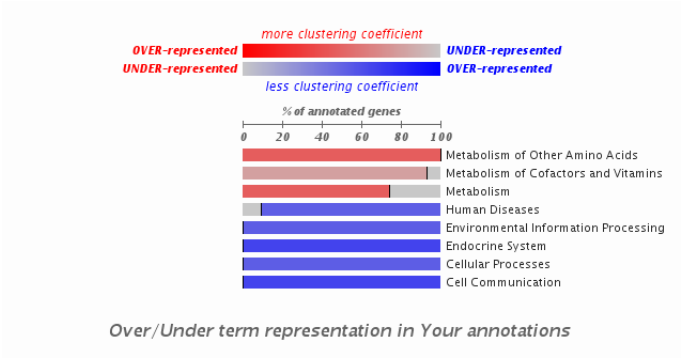


Figure 12.7: FatiScan results, clustering coefficient.

Amino Acid Metabolism, *Metabolism*, *Xenobiotics Biodegradation and Metabolism*, *Metabolism of Cofactors and Vitamins*, *Glycan Biosynthesis and Metabolism* and *Genetic Information Processing*. In this category, we observe that all terms are associated to several types of metabolism and production of proteins, both essential parts of the housekeeping tasks of the cell.

The results for betweenness centrality (Figure 12.6) are very similar to the ones obtained for connections degree. It seems that there are not classes associated to very central nodes, we have classes under-represented in external nodes though. They are *Endocrine system*, *Cell Communication* and *Cellular Processes*. Furthermore, we find one class over-represented in external nodes, *Metabolism*. Again, communication is associated to high levels of the network parameter, in this case centrality, and metabolism to low levels.

This observation is reverted when we consider the clustering coefficient (figure 12.7). Yet again there are no classes over-represented for the nodes with high level of the parameter but under-represented in the bottom of the lists (KEGGs in less connected areas). They are *Metabolism of Other Amino Acids*, *Metabolism of Cofactors and Vitamins* and *Metabolism*. The classes associated to the nodes in less connected areas are *Human Diseases*, *Environmental Information Processing*, *Endocrine System*, *Cellular Processes* and *Cell Communication*. From these observations we can say that, although there are neither hubs nor central KEGGs related to metabolism, these nodes seem to be in quite well interconnected areas where all the nodes play similar roles. Conversely, signalling related KEGGs have a completely different role in the network, they are very connected and very central nodes but the areas in which they are located do not have an interconnected neighbourhood as it is characteristic of hubs.

The analysis of the FatiScan results for the ranked list of KEGGs according to their auto-interactions, from more to less auto-connected, are similar to the results for connections degree and betweenness. The classes associated to high levels of the parameter are those related to signalling, while metabolism related classes are over-represented in low levels of the parameter (figure 12.8). This fact suggests a similar situation in KEGGs network and in the ppi network where the

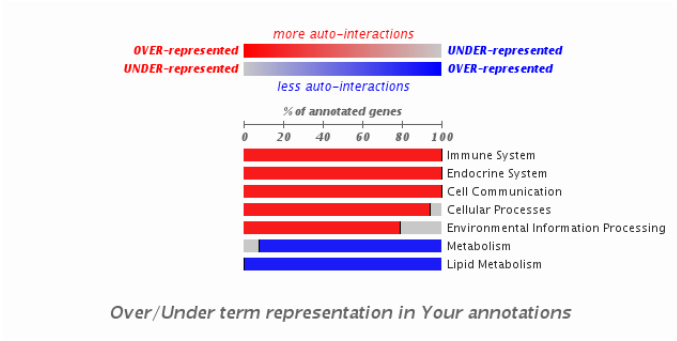


Figure 12.8: FatiScan results, auto-interactions.

more connected nodes would be proteins associated to signalling, obviously this result is biologically meaningful.

12.2 KEGGs networks in normal and cancer cellular stages

Previous sections refers to the KEGGs network generated with the complete set of ppis in the curated interactome. Even though we certainly may extract important clues about the global organization of the KEGG pathways in the cellular machinery from this analysis, it is crucial to bear in mind that the complete network never happen in the cell at the same time, principally because not all the proteins are present in all the cell types and stages.

The aim of this study was to explore the functional changes in cell phenotypes of normal and cancer stages by exploring their differences in the corresponding KEGGs networks organization. We include inter and intra nodes (KEGGs) analyses. We used a collection of SAGE libraries representing a wide spectrum of transcriptomic experiments to filter the data of the complete human interactome and generate tissue and histology specific ppi interactomes. The KEGG pathways were superimposed into this set of ppi networks to obtain a final set of 316 different KEGGs networks annotated with their tissue and histology (cancer or normal), see Materials and Methods section 8.1 for more information about this collection. See figure 8.1 for an schema of the generation of the KEGGs networks collection.

12.2.1 Normal and cancer libraries quality comparison

The first issue to address was to compare normal and cancer libraries in terms of number of tags and transcripts in order to discard that differences we report are due to differences in size or quality between normal and cancer libraries. Figure 12.9 shows the number of transcripts versus number of tags of each set of

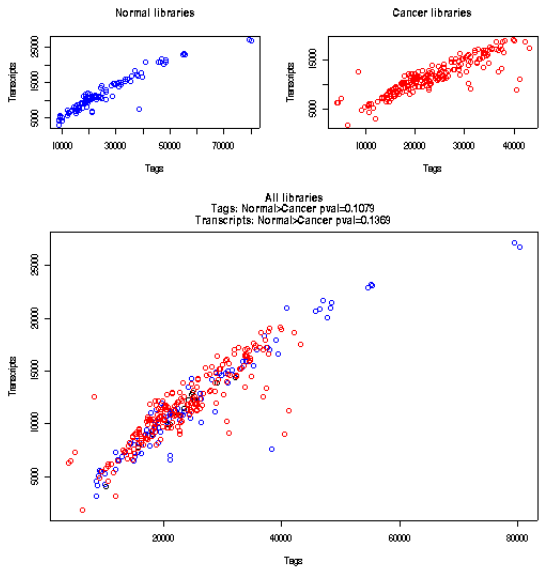


Figure 12.9: Libraries quality control.

libraries, normal and cancer, separately and a joint plot of both sets. We applied a kolmogorov-Smirnov test to compare the tags and transcripts distributions for normal and cancer sets of libraries. The p-values associated to the tests were no significant (see figure 12.9) meaning that there are no differences between normal and cancer libraries in terms of number of tags and number of transcripts.

12.2.2 Inter-KEGGs interactions variation in normal and cancer tissues

We have conceived a systematic analysis to explore the changes in the number of interactions (ppis) that connect every pair of nodes when comparing normal versus cancer KEGGs networks of the same tissue. We use the KEGGs networks inferred from the transcriptomics experiments that account for gene activity in different tissues and histologies, previously described. The changes in the number of connections in the ppi network represented within every KEGG were also considered as they are auto-connections. To measure the differences between edges' weights over two sets of libraries (cancer and normal) we developed an index: the Connectivity Index (CI), see Materials and Methods section 8.3 for details.

The CIs were calculated for every pair of KEGGs in a total of 21 tissues. Each tissue had at least one normal and one cancer KEGGs network. We generated five matrices per every tissue corresponding to the meta-categories in which KEGGs are generally classified: Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes and Human Diseases.

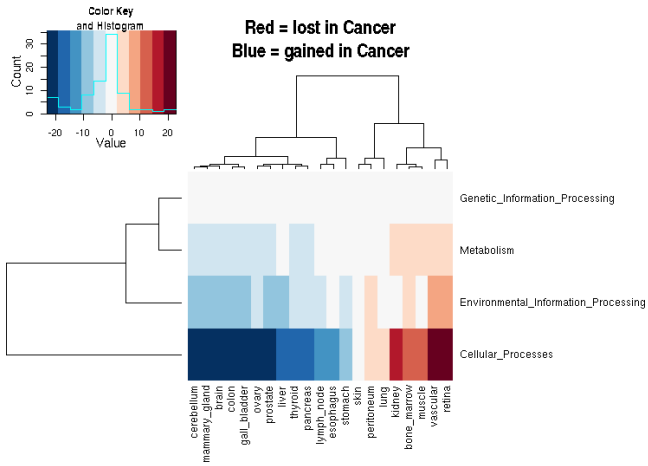


Figure 12.10: Global trend of KEGGs inter-connectivity variation from normal to cancer in different tissues. Results are broken down into general cell activities defined by KEGG database meta-categories represented in rows.Human Diseases KEGG metacategory has been removed from the graph because it does not represent general cell activities potentially altered by cancer.

Each matrix represents the differences in inter-connectivity for pairs of KEGGs given by their CI.

We generated a global picture that includes all the information from the collection of matrices about the variation in activity from normal to cancer stages in the general cell activities described by the KEGG meta-categories (figure 12.10). This graph shows how every tissue gain or lose physical connections in each of the cell activities in a categorical way (see Materials and methods section 8.3 for a detailed explanation on its generation). Bluish colors mean a global gain of connections from normal to cancer stages while reddish indicate a loss of connections. The higher the intensities, the bigger the global change in connections. Summarizing, our results indicate that Genetic Information Processing KEGG pathways do not suffer general changes in their inter-connectivity. Then, in increasing trend of change in the number of connections we have Metabolism, Environmental Information Processing and Cellular Processes. Interestingly, we observe that tissues show same trend in all the cell activities although with different level of variation. Skin does not suffer general changes in any KEGG meta-category as described by the categorical Connectivity Index. In the same terms, peritoneum, lung, kidney, bone marrow, muscle, vascular and retina show a general loss of connections in cancer. On the contrary, cerebellum, mammary gland, brain, colon, gall bladder, ovary, prostate, liver, thyroid, pancreas, lymph node, esophagus and stomach present a global gain of connections. However, these results represent global trends. A closer look into the details of each KEGG meta-category shows a less homogeneous behaviour in the set of KEGG pairs conforming the meta-categories in several tissues so a global index close to zero does not neccessary means absence of changes but it could also indicate very little changes in comparison with the rest of the analyses or even an equal variation in both sides.

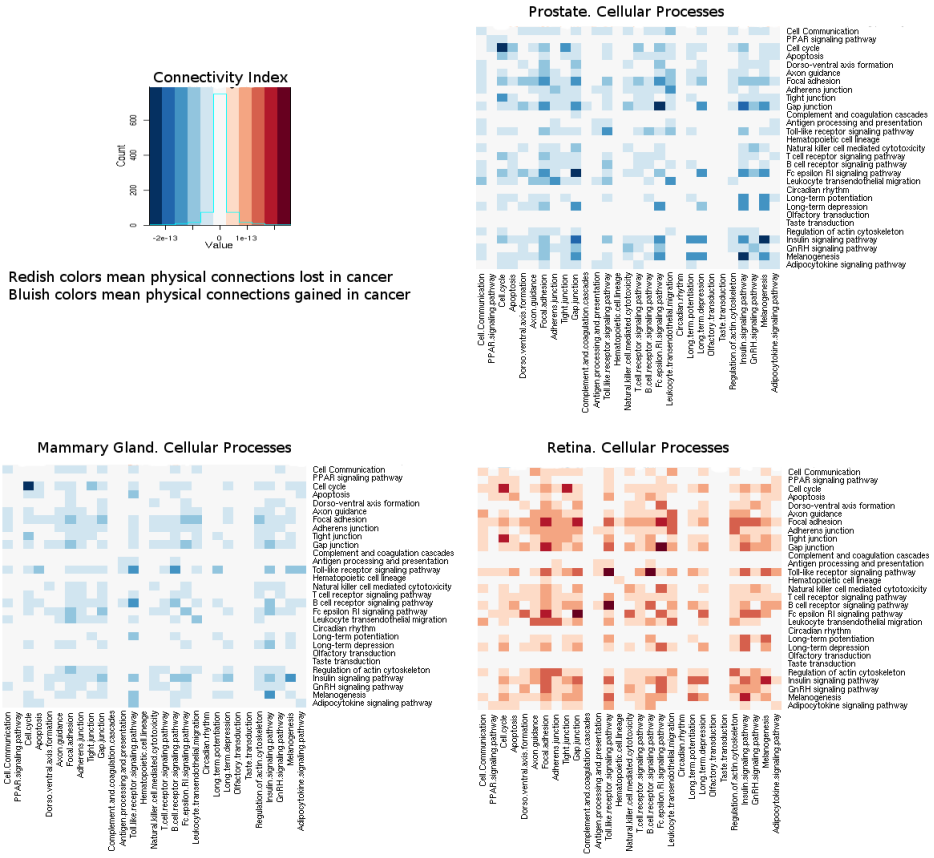


Figure 12.11: Variation of physical connections from normal to cancer stages in KEGGs pairs belonging to Cellular Processes meta-category in several tissues as measure by the Connectivity Index.

Having a more detailed observation to the KEGG pairs' CIs in tissues and KEGG meta-categories located in the extremes of the figure 12.10 we find some interesting patterns that highlight the more important pathways in the transition from normal to cancer cell stages. In the case of Cellular Processes related pathways, concerning the tissues that generally gain connections in cancer (cerebellum, brain, colon, gall bladder, mammary gland and prostate) the most important KEGG pairs seem to be: auto-connections in *Cell cycle*, *Cell cycle - Tight junction*, *Gap junction - Insulin signaling pathway*, *Gap junction - Fc epsilon RI signaling pathway*, *Toll like receptor signaling pathway* auto-connections, *Toll like receptor signaling pathway - B cell receptor signaling pathway* and *Insulin signaling pathway - Melanogenesis*, see figure 12.11 for more details. The reverse pattern appear in the tissues that generally loss connections in cancer, see figure 12.11.

The second global process that appears to be more affected is the Environ-

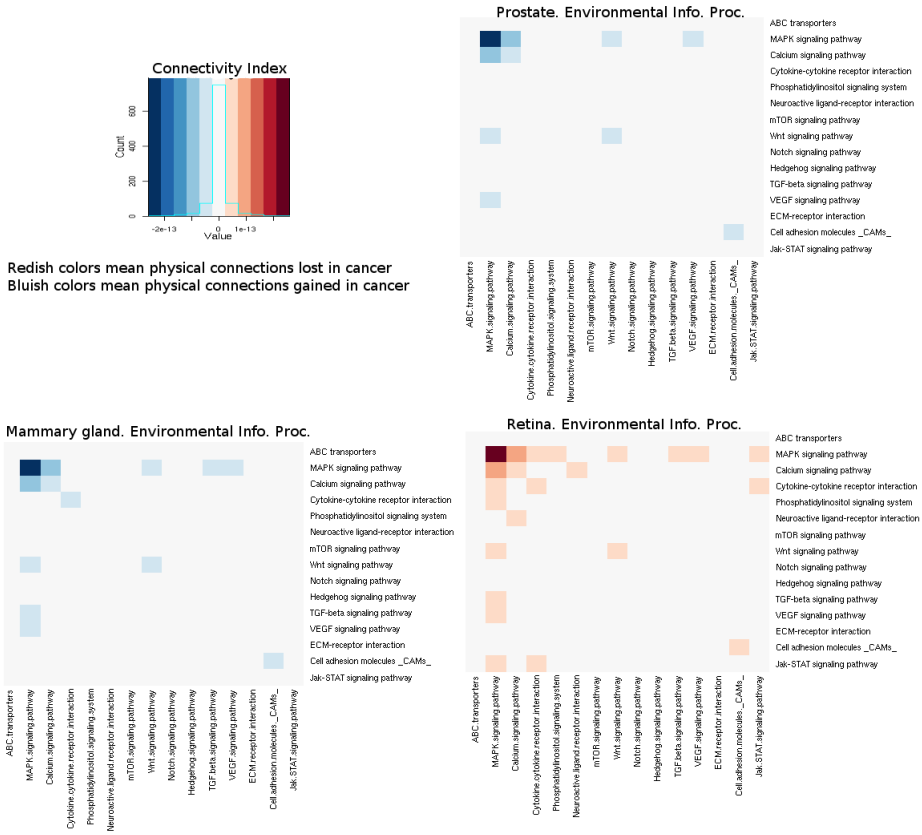


Figure 12.12: Variation of physical connections from normal to cancer stages in KEGGs pairs belonging to Environmental Information Processing meta-category in several tissues as measure by the Connectivity Index.

mental Information Processing. As figure 12.10 shows, the tissues follow the same pattern as in Cellular Processes being vascular and retina the ones that lose more connections in cancer and cerebellum, mammary gland, brain, colon, gall bladder and prostate, the ones that gain more connections. Figure 12.12 shows the KEGG pairs relations in terms of their Connectivity Indexes. Clearly the *MAPK signaling pathway* auto-connectivity is the most affected pair in both set of tissues followed by the pairs: *MAPK signaling pathway - Calcium signaling pathway*, *MAPK signaling pathway - Wnt signaling pathway*, *MAPK signaling pathway - VEGF signaling pathway*, *Wnt signaling pathway* auto-connections and *Cell adhesion molecules CAMs* auto-connections.

12.2.3 Global patterns of network features in cell phenotypes

In addition to the analysis of individual tissues and cancers, a not less interesting subject is the study of the KEGGs networks in terms of the characterization of their topology features to finally extract global patterns in the cell phenotypes introduced in the analysis.

We calculated some network features of the 316 KEGGs networks including:

- number of nodes
- number of edges
- total of edges' weights
- the sum of all the ppis in the network
- clustering coefficient mean
- number of components
- number of bicomponents

These features may characterize globally a biological network in terms of: size (number of nodes), connectivity (number of edges), activity (total edges weights), inter-connectivity meaning globally connected instead of being just a few nodes which make the connections (clustering coefficient mean) and unity (number of components and bicomponents).

We have information about several cell phenotypes including several SAGE libraries per each tissue and histology (cancer or normal) representing an excellent and diverse raw material to categorize both tissues and cancers accordingly to the relationships among the biochemical pathways acting in terms of physical interactions between their components (proteins).

We sorted all the SAGE libraries based on each of the graph parameters calculated. To avoid that the size of the library affects to the KEGGs network in its graph parameters, we divided every parameter measure by the number of transcripts in each library. The ranked lists of libraries were used as input for the FatiScan algorithm to test whether there are meta-classes associated to the distribution of the parameter (see Materials and Methods section 6.3.1 for method explanation). The meta-classes used to annotate the libraries were: tissue (e.g. brain), histology (e.g. cancer) and tissue plus histology (e.g. brain_cancer). Thus, we are going to be able to extract global patterns of network features in the set of cell phenotypes. Only a selection of FatiScan results is shown.

Figures 12.13 and 12.14 show the FatiScan results for the sorted lists of KEGGs networks according to the number of nodes and edges respectively. Number of nodes represents the size of the network while number of edges represents its

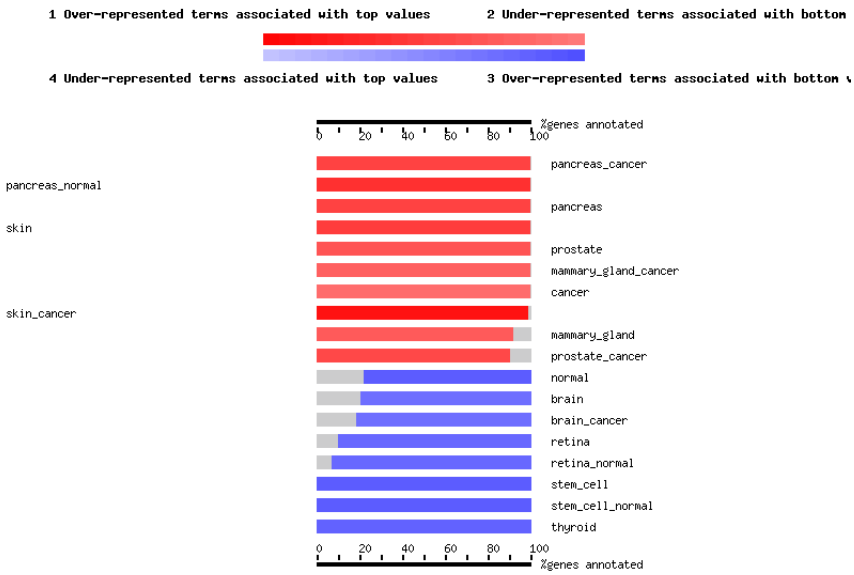


Figure 12.13: FatiScan results for lists of KEGGs networks sorted by the number of nodes (normalized).

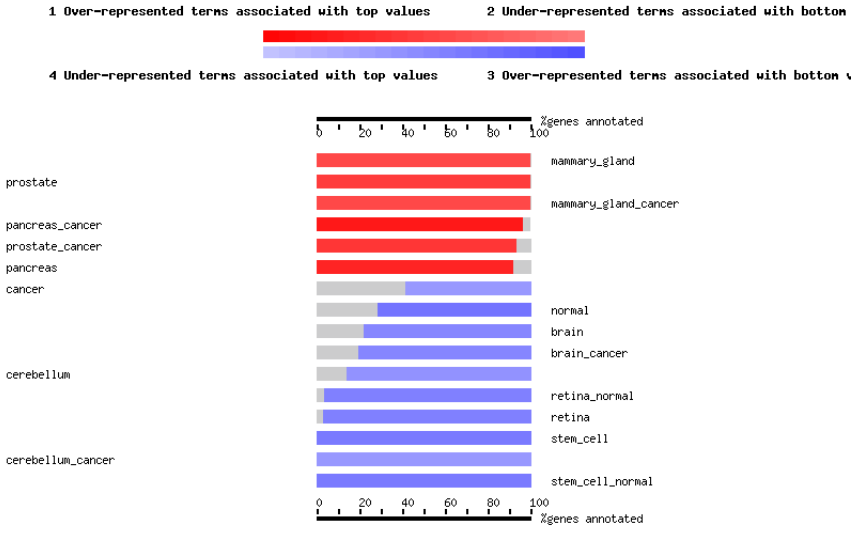


Figure 12.14: FatiScan results for lists of KEGGs networks sorted by the number of edges (normalized).

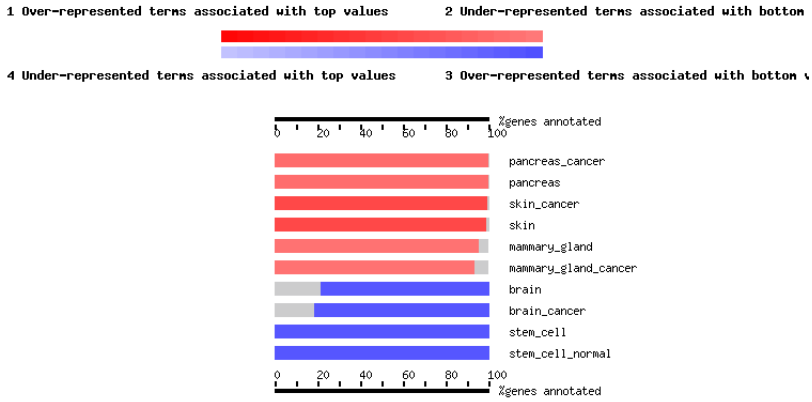


Figure 12.15: FatiScan results for lists of KEGGs networks sorted by the clustering coefficient mean (normalized).

degree of connectivity. We can see that the biggest networks are associated to *pancreas_normal*, *skin* and *skin_cancer*,. Additionally there are other set of terms that are under-represented in the bottom of the list, so still associated although more weakly to big networks, they are *pancreas_cancer*, *pancreas*, *prostate*, *mammary_gland_cancer*, *cancer*, *mammary_gland* and *prostate_cancer*. Apart from *pancreas_normal*, *skin*, *skin_cancer* and *cancer*, the rest of the terms also appear associated to networks very connected (figure 12.14). The terms associated to smallest networks are *normal*, *brain*, *brain_cancer*, *retina*, *retina_normal*, *stem_cell*, *stem_cell_normal* and *thyroid*. All of them but *thyroid* are associated to less connected networks. The joint study of both graphs show a high correlation between size (number of nodes) and connectivity (number of edges), an interesting case is *pancreas*, whereas *pancreas_normal* annotated libraries have big size, more than its cancer counterpart, it seems to be less connected than *pancreas_cancer*.

Figure 12.15 shows the FatiScan results for the lists of KEGGs networks ranked by their clustering coefficient mean. A higher clustering coefficient's mean indicates a more robust network in terms of inter-connectivity, that is, networks in which a high proportion of nodes are very connected among them. The figure 12.15 shows that there are no terms associated to highly inter-connected networks but there certainly are some terms under-represented in networks with low inter-connectivity. Therefore, although with a weaker support, we may consider the terms *pancreas_cancer*, *pancreas*, *skin_cancer*, *skin*, *mammary_gland* and *mammary_gland_cancer* as annotating high inter-connected KEGGs networks. In the other hand the terms directly associated to not inter-connected networks are *brain*, *brain_cancer*, *stem_cell* and *stem_cell_normal*.

Summarising, *pancreas_cancer*, *skin_cancer* and *mammary_gland_cancer* represent the patterns for biggest (with more connections and at the same time

most inter-connected) libraries in terms of the networks that their biochemical pathways form considering the links as the physical interactions between their proteins. *Pancreas_normal* represents a rare example of a big network not specially well connected. The libraries annotated with *brain_cancer* and *stem_cell_normal* represent the smallest, less connected and less robust KEGG networks.

Part V

Dicussion

Chapter 13

Resources for functional profiling

13.1 Tools for functional profiling

Although some of the components of Babelomics have been working since 2003, e.g. the FatiGO algorithm was first published in 2004 (Al-Shahrour *et al.*, 2004) the idea of assembling different methods (functional enrichment and gene set enrichment methods) with a large number of functional module definitions crystallized as the Babelomics project, which was first published in 2005 (Al-Shahrour *et al.*, 2005).

The availability of different flavours of functional enrichment and gene set enrichment methods that use gene modules of different nature makes of Babelomics a unique tool among other available resources of similar characteristics. Babelomics is among the most widely used web tools for functional profiling. Table 13.1 shows the number of Google Scholar entries for the top ones, as a measure of the impact of each tool in scientific community. Only four tools surpass the threshold of 500 citations and Babelomics is placed in third position of the overall ranking being the only one among those four that provides both functional enrichment and gene set analysis. If we consider the whole list, there are just another tool that provides both types of analyses, FuncAssociate. The list was generated the 25^{th} of August of 2008, obviously, any citation index is affected by the date in which the paper was pubished. Consequently, GSA methods, which are newer, are affected by this fact. During 2007 Babelomics has registered an average of 200 experiments analysed per day.

Tool	URL	Analysis type[a]	References	Citations[b]
GSEA	http:/www.broud.mit.edu/gsea/	GSA	Mootha *et al.*, 2003; Subramanian *et al.*, 2005	1391
DAVID	http://www.DAVID.niaid.nih.gov	FE	Dennis *et al.*, 2003	715
Babelomics	http://www.babelomics.org	FE, GSA	Al-Shahrour *et al.*, 2004, 2005, 2006, 2007, 2008	575
GOMiner	http://discover.nci.nih.gov/gominer/	FE	Zeeberg *et al.*, 2003, 2005	527
MAPPFinder	http://www.GenMAPP.org	FE	Doniger *et al.*, 2003	442
Ontotools	http://vortex.cs.wayne.edu/ontoexpress/	FE	Draghici *et al.*, 2003b; Khatri *et al.*, 2004, 2005, 2006, 2007	297
GOStat	http://gostat.wehi.edu.au/	FE	Beissbarth *et al.*, 2004	274
GOTM	http://genereg.ornl.gov/gotm/	FE	Zhang *et al.*, 2004	196
FunSpec	http://funspec.med.utoronto.ca	FE	Robinson *et al.*, 2002	120
FuncAssociate	http://llama.med.harvard.edu/Software.html	FE, GSA	Berriz *et al.*, 2003	111
GeneMerge	http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html	FE	Castillo-Davis and Hartl, 2003	110
GOToolBox	http://gin.univ-mrs.fr/GOToolBox	FE	Martin *et al.*, 2004	85
WebGestalt	http://bioinfo.vanderbilt.edu/webgestalt/	FE	Zhang *et al.*, 2005	71
GFINDer	http://www.medinfopoli.polimi.it/GFINDer/	FE	Masseroli *et al.*, 2004, 2005	58
PLAGE	http://dulci.biostat.duke.edu/pathways/	GSA	Tomfohr *et al.*, 2005	31
PathwayExplorer	https://pathwayexplorer.genome.tugraz.at/	FE	Mlecnik *et al.*, 2005	29
GOAL	http://microarrays.unife.it	GSA	Volinia *et al.*, 2004	29

Table 13.1: Functional profiling data analysis web tools. [a]Type of analysis: FE, functional enrichment, GSA, gene set analysis. [b]Citations are taken from Scholar Google as of 25^{th} of August 2008.

13.2 Resources to apply ppi data into Functional Genomics

Historically, in the field of ppis, the majority of the resources available have been mainly focused on visualization aspects rather than in the proper analytical steps. Most of the databases have their own visualization tools that exceptionally provide some applications to carry out very simple analyses. DIP through its satellite project LiveDIP has a tool for finding the path between two proteins. This tool also performs a simple mapping of the proteins selected in microarray experiments onto the interactome. BIND can export directly to Cytoscape (Shannon *et al.*, 2003), a popular visualization tool, and may perform some simple analysis of enrichment in GO clusters called OntoGlyphs. HPRD does not provide visualization although it does offer transcriptome information. MINT estimates the Minimal Connected Network (MCN, see Materials and Methods section 6.4.3 for complete explanation) and has a java environment available for visualization purposes. The IntAct database includes an application called MINE that can calculate and represent the MCN. BioGRID does not provide any visualization yet although the application Osprey (Breitkreutz *et al.*, 2003) uses it as underlying support database and consequently can be used as an interface to BioGRID. The database STRING (von Mering *et al.*, 2007) includes an static visualization tool with a complete set of options including generating the MCN.

Besides the facilities provided by the databases there are programs, such as Osprey (Breitkreutz *et al.*, 2003), Cytoscape (Shannon *et al.*, 2003), VisANT (Hu *et al.*, 2007) and PATIKA (Dogrusoz *et al.*, 2006), that aim to provide a general framework for ppi data management. Cytoscape and VisANT allows the development of plug-ins that can be integrated into them to perform more specific tasks. Cytoscape is probably the most successful application in this field and it has an ample community of users and developers. At the time of writing this thesis there were 48 plug-ins available. A good review about visualization and network management packages can be found in Suderman *et al.* (2007).

Other applications like the Agile Protein Interaction DataAnalyzer (APID) (Prieto *et al.*, 2006), Genes2Networks (Berger *et al.*, 2007) and PIANA (Aragues *et al.*, 2006) were developed with the aim of become a common repository for different ppi datasets. APID and Genes2Networks are web-based tools that make the datasets available. PIANA is more orientated to computer scientists as a working framework for ppi data management. It also may predict novel interactions and calculate network topological parameters.

In a more general functional profiling context, ppi data has quite recently been introduced into suites of programs like Babelomics (Al-Shahrour *et al.*, 2006, 2007, 2008) and DAVID (Dennis *et al.*, 2003) although with different grade of sophistication. DAVID simply reports the interactions associated to the genes of a list and performs a classical enrichment analysis for each of the interactions. Babelomics includes a new module called SNOW (introduced in this thesis in Results 9.4) that calculates the MCN and evaluates the significance of its robustness as functional class comparing its topological parameters versus distributions

of same sized lists of random genes or proteins. It also evaluates the presence of hubs, central nodes and highly connected areas in the pre-selected genes or proteins versus a curated interactome.

Thus, we have seen how ppi data has been used mostly as exploratory resource, however SNOW goes one step beyond than the rest of available tools reported before, apart from building the module of action using the ppi data, this is evaluated by comparing its structure with a background, the result is accompanied by a p-value showing the statistical significance of the comparison. Just as in a typical functional enrichment analysis dealing with discrete and flat annotation labels (GO terms, KEGG pathways, etc.), the comparison with the background is essential to demonstrate the relevance of the discovery. This, together with the functional informations that SNOW reports about the MCN and the interactive visualization applet make SNOW as a very valuable functional profiling tool that used under GEPAS and Babelomics environment could lead to novel discoveries in functional modules of protein.

The introduction of more structured data as the network concept in functional analysis is more and more required. Analysis of regulatory (Yeger-Lotem *et al.*, 2004), co-expression (Ghazalpou *et al.*, 2006) and genetic (Kelley & Ideker, 2005) networks are some examples of the applicability of graph theory to biological data.

Chapter 14

Gene/protein annotation for functional genomics

14.1 New sources of annotation

To perform a functional profiling analysis of a genome-scale experiment we need to pre-define the functional modules that we are going to look for, or at least, the raw data that is going to be used to build them, as in the case of ppi data. At this point we do not have a complete collection of functional modules acting in every cell type in each cell stage. In fact, that seems to be an objective far from being reachable due to the variability of cell types and possible internal and external disturbances. There are several resources that collect gene signatures able to classify cell stages by themselves. A successful instance is Oncomine (Rhodes *et al.*, 2007), where a broad set of cancer experiments are functionally analysed. L2L (Newman and Weiner, 2005) and LOLA (Cahan *et al.*, 2005) are other examples. But despite these and other efforts made in the field, it is strictly necessary to use approximations to this functional modules if we want to extract the best from the functional profiling methods. This approximations are the called functional classes and are usually defined by the functional annotation of the biological sequences. The hypothesis assumed when performing a functional profiling analysis using GO terms, KEGG pathways, etc. is that they co-express (Lee *et al.*, 2004) although early attempts to deduce gene functionality (that is, functional category membership) from gene co-expression revealed that many functional categories did not even show a detectable degree of internal co-expression (Brown *et al.*, 2000; Mateos *et al.*, 2002).

Classical functional profiling analysis usually use standard annotation to circumvent this lack of gene signatures. The most popular functional labels are probably GO terms, KEGG or BioCarta pathways and MeSH words. However, the coverage of the annotation provided by these standard functional labels is far from being complete. Moreover, genes may have multiple annotations due to both multiple function and different scope of the annotation. In this thesis we

have introduced new sources of annotation into the functional profiling analysis that may cover some percentage of unannotated genes and that supplies a complementary source of knowledge to the information already given by the more standard functional labels.

14.2 Curation of protein-protein interactions data

Obtaining a curated set of ppis as complete as possible to work with is not a trivial task. The databases' coverage, the depth and type of annotation and the lack of accuracy of some of the techniques are a limitation that have to be solved. Reguly *et al.* (2006) established a milestone in this field generating a possibly complete yeast interactome free from false positives via manual curation. For the rest of the species we are far from having this kind of detail because there is still a low coverage of the complete interactome basically due to its size. While there is not a big difference in the genome size between not very related species, it seems that in terms of ppis the differences increase drastically, suggesting that the complexity in the organisms must have a strong post-transcriptional component. For instance, human has 21,541 protein coding genes and 650,000 predicted interactions while *Caenorhabditis elegans* has 20,140 genes and three times less predicted interactions (genes count taken from ensembl genome browser release 49 and interactome size predictions taken from Stumpf *et al.*, 2007).

There is a clear necessity of methodologies to filter ppis given that manual curation is not always a feasible task. Therefore, several approaches have been proposed (review on Badet *et al.*, 2004). We will point out some of them:

- Promiscuity (Uetz *et al.*, 2000; Ito *et al.*, 2001; Gavin *et al.*, 2002; Ho *et al.*, 2002), that consists in removing proteins having many interaction partners, the called sticky proteins.
- Topological criteria (Bader & Hogue, 2002), it is specific to Co-IP experiments, retaining the bait-hit (spoke) rather than the bait-hit and hit-hit (matrix) interactions.
- Intersection of multiple high-throughput datasets (von Mering *et al.*, 2002; Deane *et al.*, 2002).
- Selecting ppis detected with two different techniques (von Mering *et al.*, 2002), combination of two methods increase coverage and accuracy.
- Intersection with other type of data, e.g. interacting proteins whose transcripts co-express are more likely to be real (Ge *et al.*, 2001; Deane *et al.*, 2002; Jansen *et al.*, 2002) or inferences of ppis across species due to protein homology (Deane *et al.*, 2002).
- Logistic regression approach (Bader *et al.*, 2004), that uses statistical and topological descriptors to predict the biological relevance of protein-protein interactions obtained from high-throughput screens.

In this thesis we have introduced a simple methodology (details in Materials and Methods 5.5.1) for ppi data curation that was applied to build our curated home-made interactome that has been used as raw data in all our analyses that use ppi data. It can be seen as a modification of the method proposed in von Mering *et al.* (2002) explained in point 3. Our method makes use of the standard annotation of the ppi data through MI ontology re-annotating the interactions to a less specific level. After that we extract the ppi annotated with at least two different annotations. By using lower levels of depth in the ontology of techniques we ensure that ppis extracted with experiments with similar basics that may have same biases in the detection process are not selected. Due to the low reliability of ppi data extracted with high-throughput techniques we consider that a good approach is to apply high restricted method for its curation.

Chapter 15

Methodologies for functional profiling

15.1 New dimensions in Functional profiling

The classical functional labels used in the field (GO terms, KEGG pathways, MeSH words) have a "flat" structure in the gene annotation. Their relationship with the biological sequences is in a *have it/don't have it* terms. In this thesis we have introduced new sources of annotation with the aim of complementing the incomplete genome annotation and to provide to the analysis different points of view that could amplify the scope of the knowledge we can extract from the analyses. However, these functional labels come with a more complex structure, which implies the development of methodologies that could deal with the particularity of each type of label. This was an essential challenge in the development of this thesis.

A requirement for this kind of methods is that they should provide a strong statistical support. A mapping of the labels into the set of genes is not enough. It must be a proper comparison with a background that assures that the enrichment is a real fact and not a common and general property of the genome.

Summarizing, in this thesis we have introduced two types of annotation. One of them are labels associated to the genes through a value (weight); this is the case of gene-bioentity co-occurrences in scientific literature and associations of gene-phenotype through a transcriptional measurement. The second type of annotation introduced are the protein-protein interactions, that are discrete labels with a supra-structure.

For each of these types of annotation, a methodology to find out the enrichment in a set of genes was developed. The set of genes to analise have the form of a list of key previously genes selected by a special feature or a ranked list of all the genes in the experiment according to a given property. Ranks or pre-selection of genes depend on the experiment design and lead to different types of algorithm

as explained in the Introduction, section 3.2.

The methods developed have been implemented also as web-tools and are part of Babelomics and GEPAS. The availability of these methodologies through these two platforms certainly assures two very important issues:

- The complete environment formed by Babelomics and GEPAS augments the capabilities of each of the modules isolated as the analysis can be performed and planned from the beginning as an easy pipeline and the results from one module can be complemented by the others.
- Babelomics and GEPAS are widely used tools with a secure prospect of maintenance.

15.2 Functional profiling applied to time-series experiments

As we learn more on the functional basis of the cooperative behaviours of groups of genes, systems biology approaches gain more importance in our attempt to decipher cell biology relevant questions. Following this, the interpretation of genome-scale experiments is starting to focus more in groups of functionally-related genes than on the properties of individual genes. Recently, different procedures that make use of functional annotations for the direct selection of functionally-related groups of genes have recently been proposed in the context of microarray data (Al-Shahrour *et al.*, 2005; Gasch *et al.*, 2000; Kim and Volsky, 2005; Mootha *et al.*, 2003; Subramanian *et al.*, 2005). These procedures use a list of genes ranked by differential gene expression and, without imposing any threshold based on the experimental values, study global over- or under-expression of blocks of functionally related genes. Nevertheless, its application has been restricted to static experimental designs. In this thesis we have shown how to expand this concept to the functional analysis of a time series. This analysis gives dynamic information on continuous behaviours occurring across the series of experiments analysed. By means of the procedure presented in this paper it is easy to understand the sequence of functional events taking place in particular moments of the period studied. It is important to remark that, for the first time, we have directly addressed the temporal evolution of the biological roles fulfilled by the genes, and not the behaviour of individual genes, which might (and actually do) contribute to more than one functional category.

As previously mentioned, the use of different GO categories (or other functional terms) allow explaining different aspects of the biology of the cell (e.g. the biological roles fulfilled by the genes by using the biological process GO category, or where these roles took place, by using the subcellular location GO category, etc.). The proposed methodology addresses systems biology-inspired questions on the behaviour of groups of functionally-related genes. The method provides results with a statistical support. The so obtained significance p-values make reference to the functional blocks of genes, but not necessarily to individual genes.

The method can be applied using any kind of gene annotation beyond the GO terms here used, such as KEGG pathways, Interpro domains, etc.

Summarizing, we have presented a method for the functional analysis of microarrays series with some dependence of autocorrelation (time series, dosage series, etc.) The method allows to recover the dynamics of the significantly over-represented functional terms along the series. A proper understanding of the biology of the cell from the perspective of systems biology need of approaches like the presented here, which tackle global functional properties cooperatively carried out by groups of genes.

15.3 Evaluate a subnetwork (module of action)

When trying to assign a functional profile to a list of genes using classical annotation such as GO terms or KEGG pathways a simple search on the annotation does not give a significant information, it has to be compared with a background to see whether the functionality found is different from the expected. There are several methodologies and resources available (reviews in Khatri and Draghici, 2005; Dopazo, 2006) that make this task quite methodical and accessible to the end user. The application of ppi data to this field is quite recent so there are not standard methodologies to be applied to the evaluation of the modules found yet.

A common approach taken is to check whether the proteins in the network are enriched in any functional category (Wilkinson and Huberman, 2004; Luo *et al.*, 2006). There are even specific tools for doing this task, BINGO (Maere *et al.*, 2005) is a java applet integrable into Cytoscape visualization tool (Shannon *et al.*, 2003) that performs GO enrichment analysis to the nodes integrating networks. Although informative this test does not guarantee in the case of negative results that the network is not a module of action due to the lack of annotation or simply because they do not share a functional category but they are indeed doing something cooperatively. In fact, functional analysis using ppis do not always overlap with label based analysis (Liu *et al.*, 2007).

Liu *et al.* (2007) proposed a systems biology orientated approach called Gene Network Enrichment Analysis (GNEA). It evaluates the association of sub-networks to a determined disease. A limitation of this method is that it must start with a set of pre-established gene signatures already associated to the disease, each of them has to have a particular annotation. The gene signatures are assembled, then the relative expression in a microarray analysis, exploring the case of study, is mapped to a global network of ppis. From the interactomic and transcriptomic information a High Scoring Matrix (HSM) is extracted as a sub-network that is highly transcriptionally affected in the disease. Finally, they evaluate the hypothesis that a particular gene signature is enriched into the sub-network. Basically, ppis in this methodology substitutes the classical differentially expression analysis but it is not taking advantage of the structured data of the biological networks.

In contrast, our approach (see Materials and Methods section 6.4.4) does

focus on the study of the network features of the modules. Indeed the topology of the biological networks has been found of crucial importance when trying to understand the role of the modules in a cellular process (Yeger-Lotem *et al.*, 2004). The methodology developed during this thesis looks for networks significantly more robust than networks generated by random lists of proteins assuring that the module evaluated is in fact a real entity acting cooperatively and not a group of proteins as any randomly selected.

15.4 Networks comparison through an heuristic approach

As described in Przulj (2006), it is very important to be able to compare two networks. From these comparisons we may extract very useful information for example about differences between healthy and disease related networks or modifications in network structure through different species to elucidate evolutionary events in the ppis graphs.

A straight forward way to address the networks comparison problem could be the full description of the networks and its direct comparison. When the networks are big this can be infeasible in computational terms (requires solving the subgraph isomorphism problem, which is an NP-complete problem). Therefore, analogous to the BLAST heuristic (Altschul *et al.*, 1990) for biological sequence comparison, we need an heuristic approach to be applied to the comparison of two networks. According to Przulj (2006) there are two different heuristic approximations to the problem:

- Global heuristics, compare the distribution of network topological parameters of two networks such as connections degree and clustering coefficient or measure the diameter of the network (average length of shortest paths).
- Local heuristics, evaluates the presence of network motifs. Several categories of motifs have been proposed (Milo *et al.*, 2002; Shen-Orr *et al.*, 2002; Milo *et al.*, 2004). The more efficient are the called graphlets, small subgraphs with topological features that occur in a network with a higher frequency than expected by random.

For large networks, a local heuristic approach seems to be more accurate (Przulj, 2006). The global approach assumes we know all the network elements and that the distributions are not influenced by external parameters. We are indeed far away from this ideal situation. For the majority of the species we do not have a complete and accurate interactome, and the techniques that have been used to their generation have several bias in the type of the ppis they detect. Thus, a direct comparison of global topological parameters may be influenced by the nature of the techniques used for the ppi detection.

Similarly to the problem of in finding an enrichment in structured network features within lists of genes/proteins, a full description of the MCN and its com-

parison with full topological features of a set of networks generated from random lists can also determine whether the list is enriched in a ppi network. Again, this approach is computationally infeasible, so we propose our own heuristic approximation to solve the problem.

Our solution was to apply a global heuristic approximation as the problems described for it in the two networks comparison does not apply to our analysis. First, we are not managing such a big networks because we are not treating with whole interactomes but with subnetworks that are activated under determined conditions. The applied methodology to generate an accurate interactome should be enough to have a high degree of confidence in the data we manage. Besides, it is important to underline the fact that the sets of lists of random nodes are generated using the same interactome as to generate the MCN, so even if a bias persists after filtering, it be would present at same level in both sides of the comparison.

So, as explained in Materials and Methods section 6.4.5 we propose to take the connections degree and the number of components in the MCN as indicative of a compact network. Thus, the two requisites we impose for a list to be considered as enriched in a ppi network are that the MCN that resumes it, has:

- A distribution of connections degree significantly greater (p-value < 0.05) than the connections degree distribution of the set of MCNs generated from same sized random lists.
- Less components than the 95% of the set of MCNs generated from same sized random lists.

The other two network features that we calculate within the MCNs but are not included in this index, betweenness and clustering coefficient, seem to be parameters more related to the shape of the network that may point out its activity, e.g. a signalling cascade network is a network with a low clustering coefficient distribution because its nodes are not in very connected areas. Nevertheless, it should indeed have more connections and fewer components than a network coming from a random list.

15.5 Network enrichment in functional classes

From the analysis of network enrichment in functional classes (Results chapter 11) we could say that ppi networks do have an important role in functionally related genes. In all the classes we found a great percentage of positive results, as described in previous section. Even in genes detected to be differentially expressed in an experiment that may not be involved in a single activity but in more than one, we could detect modules of action using ppi data. Therefore the applicability of interactomics as a source of annotation in functional profiling is more than justifiable.

An important observation is that the introduction of a non-listed node in the MCN results into a better performance in the evaluation of all the network parameters of MCNs. It is quite predictable that when introducing non-listed nodes, more shortest paths are found and consequently the resulting network will have more nodes and edges, so it will become more robust. Nevertheless, this situation should have affected in the same grade to lists of proteins sharing a GO term and to the random lists, being not the case, this indicates that the fact of introducing non-listed nodes to the paths increase dramatically the robustness of functional modules defined by ppi data pointing to the relevance of this type of proteins. Indeed, not in GO terms but in transcriptomics studies, proteins no pre-selected by expression profiling has been reported to be related to disease due to its inclusion into a network of ppis (Xu and Li, 2006; Liu *et al.*, 2007; Chuang *et al.*, 2007).

15.6 Integromics. The KEGGs networks example

In the introduction we framed this thesis within an ethereal discipline called integromics. In fact, we have introduced several types annotation for biological sequences and methodologies that can use them for the functional profile of genome-scale experiments and the aim is to provide a wide range of functional modules definitions to augment the discovery possibilities.

A clear example on an integromics experiment is the analysis of the KEGG pathways physical interconnections over a collection of tissues in normal and cancer stages (chapter 12 in Results). In this analysis we introduced three types of data, transcriptomics, protein-protein interactions and biochemical pathways to capture the variation, if any, that occurs in the transition from normal to cancer stages in both the connections between biochemical pathways and the activity of the pathways isolated.

A typical functional profiling analysis would have consisted in a differential expression analysis to find out the differences in gene expression between both normal and cancer stages and after that we would have extracted the functional classes (e.g. KEGG pathways) that are over/under-represented in cancer and normal samples. Our analysis adds two new dimensions: it reports the variation in activity of every module of action (KEGG pathways) and the changes in the connection between them. But still it goes far off from this, because it gives a measure to all the variations. This important feature makes the study an unprecedented massive analysis of the differences in functional modules activity in normal and cancer stages.

Chapter 16

Future trends

16.1 Problems with ppi data

Along this thesis we have been talking about the problems that ppi data extracted by high-throughput techniques have in accuracy and coverage. Furthermore, we have highlighted the possibilities of introducing the concept of network into the functional profiling of genome-scale experiments. In this sense, we have identified at least three main challenges that have to be approached to be able to obtain whole capabilities from this kind of data:

- High-throughput techniques produce a high proportion of false positives. Besides, there is still a low coverage of the interactome for the majority of the species. For filtering ppis according to their accuracy, literature curation does not seem to be a realistic approach, so new methodologies have to be proposed while the techniques do not overcome this limitation.
- There is a clear necessity of applying the standard annotation developed by the HUPO for the ppi experiments. This should be enough to encourage the community to take a policy of sharing data to be able to have one or several repositories with all the available ppis.
- The network nature of the functional classes that the ppis form requires more complex methodologies to study modules enrichment. The topology of the networks should be taken into account as an important parameter of the module. A protein in a network cannot be annotated just as part of the network but as a node with a special position that affects the rest of of the nodes. Moreover, the global shape of the network is characteristic of its functional activity.

Recently, an ambitious initiative has been proposed by FEBS Letters journal (Ceol *et al.*, 2008), which consists on linking scientific manuscripts with protein interactions databases through a structured summary with controlled vocabulary

that has to be filled by the authors. These kinds of approaches are going to be crucial in the quality and accessibility of biological data.

16.2 Increasing the resolution of ppi data

The methodologies developed to deal with ppi data extracted by high-throughput (reviewed in Bader *et al.*, 2004) aim just to increase accuracy and coverage, not the depth of the annotations. As said in the introduction, systems biology is the field of science that attempts to understand and model the cell behaviour considering the cell components as an intricate network of interactions and not as isolated units of action. Certainly, high-throughput techniques are making a valuable contribution to the compilation of the elements and even about their relationships (e.g. yeast two hybrid assays). However, to fully understand cellular machinery and to be able to generate models that predict the system behaviour under determined conditions, we need to go further in the details.

Post-translational modification is the major mechanism by which protein function is regulated in eukaryotes. A high quality annotated interactome with information about the phosphorilation, N- and O-glycosilations, ubiquitylations, methylations or acetylation events, just to mention some of them, would enormously increase the interactome resolution concerning information and augment its possibilities in terms of function predictive tool. However, there are non public resources that provide high quality annotation of protein interactions at the level of post-translational events for a wide range of species. The major source of this kind of information is, not surprisingly, the scientific literature, still the biggest encyclopaedia about functional genomics at the expense of general biological databases and specific ppi databases. In fact, at least in *Saccharomyces cerevisiae* the ppi datasets extracted using high-throughput techniques was found to cover only about 14% of the interactions in the literature (Reguly *et al.*, 2006). There are some limited attempts though, Reactome (Vastrik *et al.*, 2007) is a very reliable resource because it is based on manual curation of scientific papers but also because of this, still with a very low coverage; or Phospho.ELM (Diella *et al.*, 2004) also manually curated, quite extent but focused on phosphorilation.

Summarising, as the majority of the data produced with massive techniques, the information related to protein-protein interactions is rapidly increasing in terms of quantity but quite slowly in terms of quality and depth of annotation. The introduction of post-translational modifications into the annotation of protein pairwise interactions would definitely increase the potential of function prediction of networks. An effort should be made to provide a systematic and ontology-based annotation of this kind of data.

Part VI

Conclusions

From the results of this thesis we can extract the following conclusions:

1. Functional Genomics needs to introduce new sources of information to complement both the biological sequences' annotation coverage and the biological knowledge parcels that are explored.

2. The methods for functional profiling of genome-scale experiments should consider the annotation data structure and the experiment design. In the functional enrichment analysis we should take into account a comparison with the background data to define statistically significant functional modules.

3. Protein-protein interaction networks conform functional modules that are also detected integrating other functional classes such as biochemical pathways or functionally related proteins.

4. Time series experiments can be studied under a systems biology perspective leading to the extraction of the dynamics of the functional modules over time.

5. The integration of several sources of annotation into a single analysis increases the possibilities for knowledge discovery. By studying the variation of the physical connectivity between biochemical pathways in normal and cancer cell stages we can measure differences in the activity of functional modules.

6. Babelomics and GEPAS are two integrated web-based suites of tools that have demonstrated their suitability to provide a complete and system-orientated analysis of microarray and other genome-wide experiments thanks to the integration of several sources of annotation.

Part VII

Appendixes

Appendix A

Resumen en castellano

Introducción

La Genómica Funcional es la rama de la Biología Molecular dedicada a describir el comportamiento celular a partir de los datos producidos por experimentos a escala genómica. Históricamente, genes y proteínas (su forma ejecutora) han sido definidos como las unidades funcionales en la célula. Este enfoque reduccionista de la Biología Molecular ha resultado en excelentes avances en el conocimiento de base de los organismos vivos a través de la identificación y descripción de los componentes responsables de los diferentes procesos celulares. A pesar de este éxito, todavía hay numerosas y fundamentales preguntas sin responder, principalmente debido a que existen muy pocos procesos que puedan ser explicados por la acción de una proteína sola. Por el contrario, las unidades de acción participantes en los procesos biológicos parecen estar formados por módulos compuestos por varias moléculas que interaccionan (Hartwell *et al.*, 1999; Barabasi and Oltvai, 2004). Esto representó una limitación para las técnicas que se venían aplicando en biología molecular, basadas en el estudio de una o pocas moléculas al mismo tiempo. Recientemente se han desarrollado técnicas de alto rendimiento capaces de reportar la acción de miles de moléculas simultáneamente. Probablemente la técnica con más éxito entre todas ellas hayan sido los Microrrays de ADN (Schena *et al.*, 1995), capaces de medir niveles de expresión de miles de genes en un mismo experimento.

El conjunto de RNA mensajeros que se expresan en una célula en unas determinadas condiciones recibe el nombre de transcriptoma. Al contrario que el genoma (información hereditaria codificada en el ADN), el transcriptoma es una entidad dinámica, sus elementos varían su presencia y cantidad dependiendo del tipo y estado de la célula. La motivación de esta tesis fue el desarrollo de metodologías que permitan extraer módulos de acción de proteínas a partir de la descripción del transcriptoma.

Parece ser por tanto que para comprender la complicada red de interacciones que sucede en la maquinaria celular es necesario estudiar tanto sus elementos

por separado como las consecuencias de su actividad conjunta. Esta propiedad emergente sólo puede ser explicada a través de una visión holística de la célula. La Biología de Sistemas es la rama de la ciencia que viene a introducir este enfoque a la biología, trata de describir el comportamiento celular en términos de la cuantificación de las interacciones entre todos los elementos presentes.

Gracias a las técnicas de alto rendimiento, la biología molecular está acumulando grandísimas cantidades de datos acerca de los elementos implicados en la actividad celular. Este tipo de datos masivos, incluidos bajo el neologismo de datos ómicos, han revertido el procedimiento habitual de proceder en biología, típicamente se solía tener mucha información sobre unos pocos elementos, por ejemplo se sabía mucho sobre pocos genes. Hoy en día y gracias a las *ómicas* tenemos muchos datos y poca información sobre ellos. Estas técnicas llevan asociada una falta de precisión además de los problemas de almacenamiento que conllevan.

La clave para aprovechar el máximo de posibilidades que estos datos nos brindan es desarrollar métodos para desechar los falsos positivos y crear sistemas de almacenamiento y anotación efectivos y controlados. Además se ha demostrado que la información obtenida a un solo nivel (genoma o proteoma, por ejemplo) no puede explicar por sí sola el comportamiento celular (Gygi *et al.*, 1999; Anderson y Seilhammer, 1997; Lee *et al.*, 1993; Hamilton y Baulcombe, 1999) por lo que una aproximación a la solución del problema es sin duda la integración de varios tipos de estos datos.

Esta tesis surgió desde el principio con la intención de ayudar a la Biología de Sistemas desarrollando metodologías capaces de integrar tantas fuentes de información como fuera posible. Siguiendo con la fiebre de la terminología *ómica*, podríamos decir que esta tesis intenta contribuir a la integrómica, todavía otra parte de la biología moderna cuyo objetivo es la integración de varias *ómicas*.

La descripción funcional de los resultados de experimentos de alto rendimiento requiere principalmente de dos elementos: anotación de genes y proteínas y metodologías capaces de extraer los procesos que definen el comportamiento celular en una situación determinada.

En cuanto a fuentes de anotación de secuencias biológicas, en esta tesis se han utilizado básicamente tres tipos:

- **Etiquetas discretas.** Son el tipo mas habitual y utilizado. Están representadas por los términos de Gene Ontology, la rutas bioquímicas de KEGG y de BioCarta. Podríamos decir que la anotación por medio de este tipo de etiquetas viene dada en términos de *anotado/no anotado*.

- **Etiquetas continuas.** Están asociadas a los genes o proteínas por medio de un valor. En esta tesis hemos utilizado dos fuentes de anotación de este tipo: co-ocurrencias entre genes y bioentidades en la literatura científica y genes asociados a diferentes fenotipos celulares por medio de un valor de expresión.

- **Etiquetas discretas con supra-estructura.** Este es el caso de las interacciones entre proteínas. Cada proteína está asociada con otras a las que

se une físicamente, el conjunto de todos los pares de interacciones forman una red de interacciones formada por nodos (proteínas) y ejes (los eventos de interacción). Al conjunto completo de interacciones entre proteínas de una célula se le llama interactoma.

Los métodos utilizados y desarrollados en esta tesis con el fin de caracterizar funcionalmente experimentos a escala genómica podrían encuadrarse en dos categorías que se detallan a continuación.

Enriquecimiento funcional en dos pasos

Una práctica muy extendida al interpretar funcionalmente este tipo de experimentos consiste en seguir dos pasos. En el primero se realiza una selección de los genes de interés, bien porque co-expresan o porque están diferencialmente expresados en la comparación entre dos muestras que representan dos diferentes estados celulares. El siguiente paso consiste en hallar el posible enriquecimiento en algún tipo de anotación comparando la distribución que tienen las etiquetas en la lista de genes seleccionados y en el resto del genoma. Una de las herramientas mas utilizadas para realizar este tipo de análisis es FatiGO (Al-Shahrour *et al.*, 2004) pero existen otras opciones (Zeeberg *et al.*, 2003; Khatri y Draghici, 2005). En esta tesis hemos desarrollado herramientas que siguen este patrón de análisis utilizando anotaciones menos convencionales que la utilizada por FatiGO, es el caso de las herramientas Marmite (Minguez *et al.*, 2007), Tissues Mining Tool y SNOW, incluidas como módulos dentro del paquete Babelomics (Al-Shahrour *et al.*, 2008).

Análisis de enriquecimiento en conjuntos de genes

Aunque muy aceptada, la metodología que tratamos en el anterior punto presenta un inconveniente en la imposición del umbral de decisión para elegir los genes importantes en el caso de problemas supervisados como por ejemplo los análisis de expresión diferencial entre dos muestras. En estos casos, debido a que realizamos miles de test estadísticos al mismo tiempo, se impone una corrección a los p-valores muy restrictiva que evita los falsos positivos pero que sacrifica muchos falsos negativos. Por ello, otra familia de métodos inspirados en la Biología de Sistemas ha surgido para cubrir estas debilidades. Este tipo de métodos llamados colectívamente *métodos carentes de umbrales* trabajan directamente sobre una lista de genes ordenada por algún parámetro. A partir de esta ordenación tratan de encontrar etiquetas biológicas cuya anotación siga una distribución no homogénea dentro de la lista. Esto indica que la etiqueta está relacionada con el parámetro de ordenación. El método más utilizado en esta categoría es el GSEA (Mootha *et al.*, 2003; Subramanian *et al.*, 2005), otro ejemplo desarrollado en nuestro departamento es FatiScan (Al-Shahrour *et al.*, 2005). Esta tesis también ha dado lugar a un método de esta familia, MarmiteScan (Minguez *et al.*, 2007) basado en el anterior pero que utiliza etiquetas continuas en lugar de discretas.

Las metodologías de las que hablamos desarrolladas en esta tesis han sido integradas en dos exitosos paquetes de herramientas web, GEPAS, que analiza experimentos de Microarray y Babelomics, que se dedica a la interpretación funcional de una más amplia gama de experimentos de alto rendimiento.

Metodologías desarrolladas en esta tesis

Sin duda gran parte de esta tesis esta dedicada a desarrollar métodos para la interpretación funcional de experimentos de alto rendimiento, en las siguientes secciones se resumen los 3 mas importantes.

Enriquecimiento en genes específicos de determinados fenotipos celulares

La fuente de anotación son los niveles de expresión de los genes en diferentes tipos celulares. Básicamente compara la distribución de los valores de expresión de dos listas de genes, típicamente una lista de genes de interés y otra representando el resto de genes del genoma. Utiliza una serie de experimentos de transcriptómica y extrae aquellos fenotipos, definidos por un tejido y una histología, en los que una de las dos listas tiene una distribución significativamente mas alta. Este método está implementado en forma de herramienta web como un módulo del paquete Babelomics bajo el nombre de Tissues Mining Tool.

Enriquecimiento funcional de bioentidades extraídas de la literatura científica

Se desarrollaron dos métodos para tratar este tipo de anotación, el primero realiza el clásico análisis funcional en dos pasos y el segundo perteneciente a la familia de métodos que no necesitan selección de genes y que realizan un análisis de enriquecimiento en conjuntos de genes, ambos están implementados en forma de herramientas web llamadas Marmite y MarmiteScan respectivamente. Marmite y MarmiteScan están incluidas en el paquete Babelomics.

La anotación estandarizada que proporcionan los términos GO o las rutas bioquímicas de KEGG, por dar un ejemplo de las mas usadas, tiene aún una cobertura baja del genoma, sin embargo existe gran cantidad de información en la literatura científica que por estar embebida en formato de texto libre es muy difícil de utilizar de forma sistemática y masiva. Nuestro objetivo fue introducir esta información dentro de la interpretación funcional de experimentos. Utilizamos las anotaciones que nos proporcionó una técnica de minería de texto, estas anotaciones están basadas en la co-ocurrencia entre genes y palabras con algún sentido en biomedicina dentro de la misma frase en los resúmenes extraídos de PubMed. Esta co-ocurrencia es evaluada con respecto a las apariciones del gen y la palabra por si sola, dando lugar a un peso por cada par de gen y bioentidad.

Las bioentidades están clasificadas en productos químicos y palabras asociadas a enfermedades. Ambas metodologías extraen las bioentidades que están

significativamente representadas en una lista de genes o en la parte alta o baja de la lista de genes ordenados, según estemos usando Marmite o MarmiteScan.

Enriquecimiento funcional utilizando datos de interacciones entre proteínas

Como ya dijimos en la introducción las interacciones entre proteínas tienen la particularidad de formar una red cuando se consideran varias a la vez. Esta red tiene unas características que no pueden ser inferidas por el estudio de sus partes de forma aislada. Se ha visto que la topología de esta red tiene implicaciones en su rol en procesos celulares (Yeger-Lotem *et al.*, 2004) por lo que su estudio podría proporcionar información relevante acerca de su funcionalidad. Existen gran cantidad de mediciones que se pueden hacer sobre una red, en esta tesis hemos elegido algunas para caracterizarla. Las que se refieren a los nodos son: el número de conexiones, la centralidad y el coeficiente de agrupamiento. Además calculamos el número de componentes que son grupos de nodos conectados entre sí y bicomponentes que son grupos de nodos conectados entre sí y a otro grupo por medio de un solo eje, este eje recibe el nombre de punto de articulación.

La peculiaridad de este tipo de análisis es que, al contrario que en los realizados con cualquiera de las anotaciones mas usadas (incluyendo las anteriores), en este caso no disponemos de clases funcionales predeterminadas que evaluar. Por lo tanto los pasos que se proponen en esta tesis son primero generar la clase funcional y después evaluarla.

La clase funcional la generamos a partir de una lista de genes que han sido seleccionados por alguna razón (co-expresión o expresión diferencial por ejemplo). A continuación se hallan los caminos más cortos que unen todos los pares de proteínas y seleccionamos aquellos que las unan directamente o a través de un número determinado de nodos no presentes en la lista (normalmente uno). La red resultante se llama Red de Conexión Mínima (RCM) y sería la red de proteínas, en función de sus interacciones físicas, que definiría la clase funcional activa bajo los criterios por los que habíamos seleccionado los genes de la lista.

Para la evaluación, nuestra metodología toma en cuenta los parámetros de la red anteriormente expuestos como indicadores de redes robustas. La distribución de estos parámetros es comparada por medio de un test estadístico con las distribuciones de un conjunto de RCMs generadas a partir de listas del mismo tamaño de genes seleccionados aleatoriamente como indicador de la señal de fondo. Para determinar si una lista está enriquecida en una red de proteínas imponemos dos condiciones:

- La distribución del grado de conexión ha de ser significativamente mayor (p-valor < 0.05) que la distribución del conjunto de RCMs generadas a partir de listas de genes cogidos al azar.
- El número de componentes de la clase funcional ha de ser menor que el 95% del conjunto de RCMs generadas a partir de listas de genes cogidos al azar.

Objetivos

Esta tesis comenzó con el objetivo general de desarrollar metodologías para la anotación funcional de experimentos a escala genómica. Más específicamente queríamos introducir nuevas fuentes de anotación que aumentaran el alcance de estos análisis complementando ambos, la cobertura de la anotación y las parcelas de conocimiento biológico que son exploradas.

Para ello requeríamos conseguir una serie de objetivos más concretos que se enumeran a continuación:

- La introducción de nuevas fuentes de información en la interpretación funcional de experimentos. Más detalladamente nos propusimos manegar tres tipos de datos:
  - Palabras con sentido en contexto biológico asociadas a genes a través de su aparición conjunta en la literatura científica.
  - Información de fenotipo asociada a los genes por medio de valores de expresión.
  - Datos de interacciones entre proteínas.
- Desarrollar metodologías para este tipo de análisis que tengan en consideración ambas, la estructura de la fuente de anotación y el diseño previo del experimento. Específicamente nos propusimos generar métodos que pudieran adaptarse a:
  - Etiquetas contínuas asociadas a los genes por medio de un valor.
  - Etiquetas simples con una supra-estructura en forma de red.
  - Experimentos que incluyan series temporales.
  - Diferentes diseños experimentales tales como problemas supervisados y no supervisados.
- La implementación de estas metodologías en herramientas web que pudieran ser integradas en los paquetes Babelomics y GEPAS diseñados para la interpretación funcional de experimentos y el análisis de microarrays respectivamente.
- Testar las posibilidades científicas de los métodos por medio de:
  - Interpretar funcionalmente experimentos de escala genómica.
  - Explorar el rol de las interacciones entre proteínas en otras clases funcionales.
  - La integración de varias fuentes de anotación para estudiar la variación de módulos funcionales en cáncer.

Resultados y Discusión

Babelomics

Esta tesis ha contribuido al desarrollo de Babelomics, un paquete de herramientas accesible vía web para la anotación funcional de experimentos de alto rendimiento. Babelomics cuenta con diferentes módulos interconectados entre sí y con otro paquete de gran difusión dentro de la comunidad científica, GEPAS, cuya funcionalidad es analizar experimentos de microarray.

Aparte de las herramientas comentadas en la metodología, Marmite, MarmiteScan y Tissues Mining Tool, la siguiente versión de Babelomics vendrá con un nuevo módulo que utiliza interacciones entre proteínas para anotar funcionalmente listas de genes. Este módulo está implementado bajo el nombre de SNOW (Studying Networks in the Omic World, *estudiando redes dentro del mundo ómico* en castellano) e integra la metodología anteriormente explicada para generar y evaluar una clase funcional formada por interacciones entre proteínas. Además SNOW evalúa estadísticamente si la lista tiene un contenido mayor de proteínas con alto grado de conexiones, más centradas o con una vecindad mas interconectada que la señal de fondo dada por el resto de proteínas del interactoma completo. La RCM es también anotada funcionalmente a través de un mapeo de los términos GO y descripciones génicas/proteicas de cada uno de sus componentes. SNOW cuenta con un sistema interactivo para visualizar la RCM que completa la funcionalidad del módulo.

Interpretación funcional de experimentos de microarray basados en series temporales

Aunque normalmente los experimentos de microarray están diseñados para estudiar condiciones estáticas, existe otro tipo en el que se miden condiciones a lo largo del tiempo. Los experimentos que incluyen una serie temporal pueden darnos información a cerca de la dinámica de la activación de genes. Las peculiaridades de este tipo de experimentos hacen que no puedan ser estudiadas bajo las mismas reglas que un experimento estático. En esta tesis diseñamos un método para analizar experimentos de microarray basados en una toma secuencial de datos a lo largo del tiempo. Este método se basa en FatiScan (Al-Shahrour *et al.*, 2006) y tiene en consideración que la medición de la expresión génica está condicionada por la medición en el tiempo anterior. Además busca módulos de genes sobre-representados en listas ordenadas de genes, es decir, está orientado de forma sistémica sin imponer umbrales de decisión.

El método fue aplicado al estudio del ciclo intra-sanguíneo de *Plasmodium falciparum*. Este ciclo es el responsable de la malaria en humanos por lo que el conocimiento de la dinámica del comportamiento celular del parásito es de crucial importancia para el desarrollo de vacunas y drogas que puedan paliar los efectos de la enfermedad. Utilizamos los términos GO para la anotación funcional. Como resultado obtuvimos la dinámica de activación y desactivación

conjunta de las funcionalidades en cada hora de las 48 que dura el ciclo completo del parásito.

Estudio del rol de las redes de interacciones entre proteínas en la anotación funcional de experimentos de alto rendimiento

Con el objetivo de descifrar el rol de las redes de interacciones entre proteínas en diferentes tipos de clases funcionales desarrollamos un análisis masivo y sistemático de las redes funcionales, de acuerdo a la descripción de estas que hicimos en la metodología, que existen dentro de cuatro tipos de clases funcionales: términos GO, rutas bioquímicas definidas por KEGG y por BioCarta y módulos de co-expresión en diferentes tipos de cáncer. Además en el análisis incluimos un conjunto de listas de genes diferencialmente expresados, ya sea inducidos o reprimidos en diferentes experimentos de microarray de una temática variada.

Utilizamos los términos GO para realizar una primera aproximación al problema. Definimos un conjunto de listas de proteínas por su anotación conjunta a un término GO. A partir de cada una de estas listas generamos una RCM que a continuación fue evaluada de la forma descrita anteriormente. Para cada lista generamos dos RCMs, una sólo con las proteínas anotadas con ese determinado GO y otra permitiendo la introducción en los caminos mínimos de una proteína externa a la lista. Los resultados demuestran que:

- Los términos GO contienen en un gran porcentaje una red de interacciones entre proteínas que es mas robusta que un conjunto de redes generadas a partir de listas sin sentido biológico.
- La introducción de una proteína externa en los caminos mínimos es traducida en que encontramos aproximadamente el doble de redes mas robustas que las generadas a partir de listas aleatorias.

Para extender estas conclusiones realizamos el mismo análisis para las listas descritas anteriormente. Las listas provenientes de experimentos de microarray fueron separadas en inducidas y reprimidas y en dos histologías (normal y cáncer) con el fin de investigar una posible diferenciación en el rol biológico de las redes de proteínas en cada una de estas situaciones.

La primera observación es que las rutas bioquímicas definidas por KEGG contienen un mayor porcentaje de redes robustas seguidas de los términos GO y a continuación por las rutas definidas por BioCarta y los módulos de co-expresión en cáncer. Las listas provenientes de experimentos de microarray resultaron las que tenían menos redes definidas con sentido biológico. Además la clasificación de estas listas en inducidas, reprimidas, normales y cancerosas no demostró apenas ninguna diferencia exceptuando en la comparación del coeficiente de agrupamiento en la que las listas anotadas como cáncer y/o reprimidas mostraron mejores resultados indicando que sus redes de proteínas podrían estar mas interconectadas que las normales y/o inducidas.

Análisis de la variación de las conexiones físicas entre rutas bioquímicas en situaciones normales y cancerosas

Diseñamos un experimento que analiza como cambian las conexiones físicas entre procesos celulares en varios tejidos entre situaciones normales y de cáncer. Los procesos biológicos están definidos por rutas bioquímicas de la base de datos KEGG.

Lo primero que hicimos fue superponer las anotaciones KEGG sobre el interactoma humano y reajustar los nodos a cada una de las rutas definidas por este tipo de anotación y los ejes a las interacciones físicas que suceden entre proteínas de dos grupos diferentes de KEGGs. Con esto conseguimos tener una red que muestra como se conectan las rutas bioquímicas en función de las interacciones físicas de sus elementos.

Para la descripción de esta nueva red utilizamos los mismos parámetros que en las redes de proteínas mas un nuevo concepto, el de pesos de los ejes representando el número de interacciones entre proteínas que se dan en cada par de nodos conectados.

Comprobamos que la red de rutas KEGG se comporta igual que una red de proteínas en cuestión de topología, ambas son redes libres de escala que son la forma mas característica de las redes biológicas. Su principal característica es que los nodos tienen en general pocas conexiones exceptuando unos pocos llamados *hubs*. Los nodos mas conectados y centrales resultaron ser los asociados a señalización mientras que los menos conectados fueron los asociados con metabolismo. En contraste, si consideramos la conectividad en el entorno cercano a los nodos, la situación se revierte.

Utilizamos datos de transcriptómica para filtrar la red de KEGGs y generar redes específicas de diversos tejidos en estados normales y cancerosos. A partir de estos datos hicimos comparaciones para cada tejido de como se ganaban o perdían interacciones físicas entre pares de KEGGs en las redes normales y cancerosas. Esta forma de analizar este tipo de situaciones toma en cuenta tres tipos de datos: de expresión, de interacción y de pertenencia a una ruta bioquímica. Además frente a un análisis clásico de enriquecimiento funcional que sólo aporta datos descriptivos, aquí tenemos mediciones de cómo varían las conexiones entre funcionalidades siendo esto un paso mas hacia la Biología de Sistemas.

Conclusiones

De los resultados de esta tesis podemos extraer las siguientes conclusiones:

1. La Genómica funcional necesita introducir nuevas fuentes de información para complementar tanto la cobertura de la anotación de secuencias biológicas como las parcelas de conocimiento en biología que son exploradas.
2. Los métodos de interpretación funcional de experimentos a escala genómica deben considerar la estructura de la fuente de anotación así como el diseño

del experimento. En un análisis de enriquecimiento funcional debemos tener en cuenta una comparación con la señal de fondo para ser capaces de definir módulos funcionales con soporte estadístico suficiente.

3. Las redes formadas por proteínas que interaccionan físicamente pueden conformar módulos funcionales que son además detectados formando parte de otras clases funcionales tales como rutas bioquímicas o proteínas funcionalmente relacionadas.

4. Los experimentos que incluyen series temporales pueden ser estudiados desde una perspectiva de sistema dando como resultado la descripción de la dinámica de los módulos funcionales en el tiempo.

5. La integración de varias fuentes de información en un solo análisis aumenta las posibilidades de extraer conocimiento. Estudiando la variación de la conectividad física entre rutas bioquímicas en estados celulares normales y cancerosos podemos llegar a medir la diferencia en la actividad de módulos funcionales.

6. Babelomics y GEPAS son dos paquetes de herramientas web que han demostrado su capacidad para integrar diferentes fuentes de información con el fin de proporcionar un análisis completo bajo una perspectiva de sistema tanto en experimentos de microarray como en otros tipos de experimentos a escala genómica.

Appendix B

Author publications

1. Minguez P, Dopazo J. (2008) **Protein Interactions for Functional Genomics** (Book chapter). Biological Data Mining in Protein Interaction Networks. IGI Global. In press.

2. Horcajadas JA, Mínguez P, Dopazo J, Esteban FJ, Domínguez F, Giudice LC, Pellicer A, Simón C. (2008) **Controlled ovarian stimulation induces a functional genomic delay of the endometrium with potential clinical implications.** J Clin Endocrinol Metab. In press.

3. Al-Shahrour F, Carbonell J, Minguez P, Goetz S, Conesa A, Tárraga J, Medina I, Alloza E, Montaner D, Dopazo J. (2008) **Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments.** Nucleic Acids Res.; 36(Web Server issue):W341-6.

4. Tárraga J, Medina I, Carbonell J, Huerta-Cepas J, Minguez P, Alloza E, Al-Shahrour F, Vegas-Azcárate S, Goetz S, Escobar P, Garcia-Garcia F, Conesa A, Montaner D, Dopazo J. (2008) **GEPAS, a web-based tool for microarray data analysis and interpretation.** Nucleic Acids Res.; 36(Web Server issue):W308-14. Epub 2008 May 28.

5. Minguez P, Al-Shahrour F, Montaner D, Dopazo J. (2007) **Functional profiling of microarray experiments using text-mining derived bioentities.** Bioinformatics.; 23(22):3098-9.

6. Al-Shahrour F, Minguez P, Tárraga J, Medina I, Alloza E, Montaner D, Dopazo J. (2007) **FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments.** Nucleic Acids Res.; 35(Web Server issue):W91-6.

7. Al-Shahrour F, Arbiza L, Dopazo H, Huerta-Cepas J, Mínguez P, Montaner D, Dopazo J. (2007) **From genes to functional classes in the study of biological systems.** BMC Bioinformatics.; 8:114.

8. Minguez P, Al-Shahrour F, Dopazo J. (2006) **A function-centric approach to the biological interpretation of microarray time-series.** Genome Inform. 2006;17(2):57-66.

9. Montaner D, Tárraga J, Huerta-Cepas J, Burguet J, Vaquerizas JM, Conde L, Minguez P, Vera J, Mukherjee S, Valls J, Pujana MA, Alloza E, Herrero J, Al-Shahrour F, Dopazo J. (2006) **Next station in microarray data analysis: GEPAS.** Nucleic Acids Res.; 34(Web Server issue):W486-91.

10. Al-Shahrour F, Minguez P, Tárraga J, Montaner D, Alloza E, Vaquerizas JM, Conde L, Blaschke C, Vera J, Dopazo J. (2006) **BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments.** Nucleic Acids Res.; 34(Web Server issue):W472-6.

11. Vaquerizas JM, Conde L, Yankilevich P, Cabezón A, Minguez P, Díaz-Uriarte R, Al-Shahrour F, Herrero J, Dopazo J. (2005) **GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data.** Nucleic Acids Res.; 33(Web Server issue):W616-20.

12. Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J. (2005) **BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments.** Nucleic Acids Res.; 33(Web Server issue):W460-4.

Part VIII

Bibliography

- Aach J and Church GM. (2001) **Aligning gene expression time series with time warping algorithms**. Bioinformatics, 17(6):495, 508.
- Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF. (1991) **Complementary DNA sequencing: expressed sequence tags and human genome project**. Science. 252(5013):1651-6
- Al-Shahrour F, Díaz-Uriarte R & Dopazo J. (2004) **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.** Bioinformatics 20: 578-580
- Al-Shahrour F, Diaz-Uriarte R, and Dopazo J. (2005) **Discovering molecular functions signicantly related to phenotypes by combining gene expression data and biological information**. Bioinformatics, 21(13):2988-2993.
- Al-Shahrour F, Minguez P, Tarraga J, Montaner D, Alloza E, Vaquerizas JM, Conde L, Blaschke C, Vera J and Dopazo J. (2006) **BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments.** Nucleic Acids Res., 34(Web Server issue):W472-W476.
- Al-Shahrour F, Minguez P, Tárraga J, Medina I, Alloza E, Montaner D & Dopazo J. (2007). **FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interpretation data with microarray experiments.** Nucleic Acids Research 35 (Web Server issue): W91-96
- Al-Shahrour F, Carbonell J, Mínguez P, Goetz S, Conesa A, Tárraga J, Medina I, Alloza E, Montaner D, Dopazo J. (2008) **Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments.** Nucleic Acids Res. 36:341-346
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) **Basic local alignment search tool.** J Mol Biol, 215(3):403-10
- Anderson L and Seilhamer J. (1997) **A comparison of selected mRNA and protein abundances in human liver.** Electrophoresis, 18, 533-537.
- Andrade MA and Valencia A. (1998) **Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families.** Bioinformatics, 14, 600–607.
- Aragues R, Jaeggi D and Oliva B. (2006) **PIANA: Protein Interactions and Network Analysis.** Bioinformatics, 22(8):1015-7
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. (2000) **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** Nat Genet, 25, 25-29.

- Badano,JL and Katsanis,N. (2002) **Beyond Mendel: an evolving view of human genetic disease transmission.** Nat. Rev. Genet., 3, 779–789.
- Bader GD & Hogue CWV (2002) **Analyzing yeast protein-protein interaction data obtained from different sources.** Nature Biotechnology 20, 991 – 997
- Bader GD, Betel D and Hogue CWV. (2003) **BIND: The Biomolecular Interaction Network Database.** Nucleic Acids Research, 31, 1
- Bader JS, Chaudhuri A, Rothberg JM, Chant J. (2004) **Gaining confidence in high-throughput protein interaction networks.** Nature Biotechnology, 22, 1, 78-84
- Bar-Joseph Z. (2004) **Analyzing time series gene expression data**. Bioinformatics, 2004, 20(16):2493-2503.
- Bar-Joseph Z, Gerber G, Simon I, Gifford DK, and Jaakkola TS. (2003) **Comparing the continuous representation of time-series expression proles to identify differentially expressed genes**. Proc. Natl. Acad. Sci. USA, 100(18):10146-10151.
- Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, and Gifford DK. (2003) **Computational discovery of gene modules and regulatory networks**. Nat. Biotechnol, 21(11):1337-1342.
- Barabasi AL, Albert R. (1999) **Emergence of scaling in random networks.** Science 286(5439):509-12.
- Barabasi AL, Bonabeu E. (2003) **Scale-Free Networks.** Scientific American 288(5):60-9
- Barabasi AL, Oltvai ZN. (2004) **Network Biology: Understanding the cell's functional organization.** Nature Reviews 5(2):101-13
- Barry WT, Nobel AB and Wright FA. (2005) **Significance analysis of functional categories in gene expression studies: a structured permutation approach.** Bioinformatics, 21, 1943-1949.
- Batada NN, Hurst LD, Tyers M. (2006) **Evolutionary and Physiological Importance of Hub Proteins.** Plos Computational Biology, 2, 7, e88.
- Beissbarth T, Speed TP. (2004) **GOstat: find statistically overrepresented Gene Ontologies within a group of genes.** Bioinformatics;20(9):1464-5.
- Ben Mamoun C, Gluzman IY, Hott C, MacMillan SK, Amarakone AS, Anderson DL, Carlton JM, Dame JB, Chakrabarti D, Martin RK, Brownstein BH and Goldberg DE. (2001) **Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite Plasmodium falciparum revealed by microarray analysis**. Mol. Microbiol, 39(1):26-36.

- Benjamini Y and Hochberg Y. (1995) **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** Journal Royal Statistical Society B, 57, 289-300.
- Benjamini Y and Yekutieli D. (2001) **The control of false discovery rate in multiple testing under dependency**. Annals of Statistics, 29:1165-1188.
- Berger SI, Posner JM, Ma'ayan A. (2007) **Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases.** BMC Bioinformatics, 8:372
- Berggard Tm Linse S, and James P. (2007) **Methods for the detection and analyisis of protein-protein interactions.** Proteomics, 7, 2833-2842
- Berriz GF, King OD, Bryant B, Sander C, Roth FP. (2003) **Characterizing gene sets with FuncAssociate.** Bioinformatics.;19(18):2502-4.
- Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, and DeRisi JL. (2003) **The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum**. PLoS Biol, 1(1):E5.
- Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bähler J, Wood V, Dolinski K, Tyers M. (2008) **The BioGRID Interaction Database: 2008 update.** Nucleic Acids Research, 36, Database issue, D637-D640
- Breitkreutz B, Stark C and Tyers M. (2003) **Osprey: a network visualization system.** Genome Biology, 4:R22
- Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr, Haussler D. (2000) **Knowledge-based analysis of microarray gene expression data by using support vector machines.** Proc Natl Acad Sci USA, 97(1):262-267.
- Bruggeman FJ and Weterhoff HV. (2006) **The nature of systems biology.** Trends in Microbiology, 15, 1, 45-50
- Brunner HG and van Driel MA. (2004) **From syndrome families to functional genomics.** Nat. Rev. Genet., 5, 545–551.
- Cahan P, Ahmad AM, Burke H, Fu S, Lai Y, Florea L, Dharker N, Kobrinski T, Kale P, McCaffrey TA. (2005) **List of lists-annotated (LOLA): a database for annotation and comparison of published microarray gene lists.** Gene.;360(1):78-82.
- Camargo A & Azuaje F. (2007) **Linking Gene Expression and Functional Network Data in Human Heart Failure.** PloS One, 12, e1347
- Castillo-Davis CI, Hartl DL. (2003) **GeneMerge–post-genomic analysis, data mining, and hypothesis testing.** Bioinformatics; 19(7):891-2.

- Ceol A, Chatr-Aryamontri A, Liacata L, Cesareni G. (2008) **Linking entries in protein interaction database to structured text: The FEBS Letters experiment.** FEBS Lettwers, 585, 1171-1177

- Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G. (2006) **MINT: the Molecular INTeraction database** Nucleic Acids Research;35(Database issue):D572-4.

- Cho S, Park SG, Lee DH, Park BC. (2003) **Protein-protein Interaction Networks: from Interactions to Networks.** Journal of Biochemistry and Molecular Biology, 37, 1, 45-52.

- Chuang H, Lee E, Liu Y, Lee, D and Ideker T. (2007) **Network-based classification of breast cancer metastasis.** Molecular Systems Biology, 3, 140

- Conesa A, Gotz, S, Garcia-Gomez JM, Terol J, Talon M and Robles M. (2005) **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** Bioinformatics, 21, 3674-3676.

- Deane CM, Salwinski L, Xenarios I, Eisenberg D. (2002) **Protein Interactions: two methods for assessment of the realiability of high throughput observations.** Mol. Cel. Proteomics, 1, 349-356.

- Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. (2003) **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** Genome Biology; 4(5):P3.

- Dijkstra, E (1959) **A note on two problems in connexion with graphs.** Numerische Mathematik, 1:269-271.

- Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR. (2003) **MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data.** Genome Biol.;4(1):R7.

- Dopazo J. (2006) **Functional interpretation of microarray experiments.** Omics, 10, 398-410.

- Dopazo J. (2008) **Formulating and testing hypotheses in functional genomics.** Artif Intell Med. In press.

- Draghici S, Khatri P, Martins RP, Ostermeier GC, and Krawetz SA. (2003) **Global functional profiling of gene expression**. Genomics, 81(2):98-104.

- Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA, Tainsky MA. (2003b) **Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate.** Nucleic Acids Res.;31(13):3775-81.

- Drewes G and Bouwmeester T. (2003) **Global approaches to protein-protein interactions.** Current Opinion in Cell Biology. 15:199-205

- Dogrusoz U, Erson EZ, Giral E, Demir E, Babur O, Cetintas A, Colak R. (2006) **PATIKAweb: a Web interface for analyzing biological pathways through advanced querying and visualization.** Bioinformatics, 22, 3, 374-375

- Eisen MB, Spellman PT, Brown PO and Botstein D. (1998) **Cluster analysis and display of genome-wide expression patterns.** Proc Natl Acad Sci U S A, 95, 14863-14868.

- Ernst J, Nau GJ, and Bar-Joseph Z. (2005) **Clustering short time series gene expression data.** Bioinformatics, 21(Suppl 1):i159-i168.

- Falk R, Ramström M, Ståhl S, Hober S. (2007) **Approaches for systematic proteome exploration.** Biomolecular Engineering 24, 155-168

- Formstecher E, Aresta S, Collura V, Hamburger A, Meil A, Trehin A, Reverdy C, Betin V, Maire S, Brun C, Jacq B, Arpin M, Bellaiche Y, Bellusci S, Benaroch P, Bornens M, Chanet R, Chavrier P, Delattre O, Doye V, Fehon R, Faye G, Galli T, Girault JA, Goud B, de Gunzburg J, Johannes L, Junier MP, Mirouse V, Mukherjee A, Papadopoulo D, Perez F, Plessis A, Rossé C, Saule S, Stoppa-Lyonnet D, Vincent A, White M, Legrain P, Wojcik J, Camonis J, Daviet L. (2005) **Protein interaction mapping: a Drosophila case study.** Genome Research, 15, 376-384

- Friedman N, Linial M, Nachman I and Pe'er D. (2000) **Using Bayesian networks to analyze expression data**. J. Comp. Biol, 7(3-4):601-620.

- Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A. (2006) **Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets.** Nat. Genet., 38, 285–293.

- Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D and Brown PO. (2000) **Genomic expression programs in the response of yeast cells to environmental changes**. Mol. Biol. Cell, 11(12):4241-4257.

- Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. (2002) **Functional organization of the yeast proteome by systematic analysis of protein complexes.** Nature, 415, 141-147

- Ge H, Walhout AJ, Vidal M. (2003) **Integrating 'omic' information: a bridge between genomics and systems biology.** Trends Genet. 19,551-560.

- Ge H, Liu Z, Church GM, Vidal M. (2001) **Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae.** Nature Genetics, 29, 482-486

- Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S. (2006) **Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight.** Plos Genetics, 2, 8, e130

- Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM. (2003) **A protein interaction map of Drosophila melagaster.** Science, 302, 1727-1736

- Girvan M and Newman MEJ. (2002) **Community structure in socila and biological networks.** Proc. Natl. Acad. Sci., 99, 12, 7821-7826

- Goeman JJ, van de Geer SA, de Kort F and van Houwelingen HC. (2004) **A global test for groups of genes: testing association with a clinical outcome**. Bioinformatics, 20(1):93-99.

- Goeman JJ, Bühlmann P. (2007) **Analyzing gene expression data in terms of gene sets: methodological issues.** Bioinformatics, 23(8):980-7.

- Götz S, Williams TD, Nagaraj SH, Nueda MJ, Terol J, García-Gómez JM, Robles M, Talón M, Dopazo J, Conesa A. (2008) **High-throughput functional annotation and data mining with the Blast2GO suite.** Nucleic Acid Research Research 36(10):3420-35

- Gygi SP, Rochon Y, Franza BR and Aebersold R. (1999) **Correlation between protein and mRNA abundance in yeast.** Mol Cell Biol, 19, 1720-1730.

- Hamilton A and Baulcombe D. (1999) **A species of small antisense RNA in posttranscriptional gene silencing in plants.** Science 286 (5441): 950–2

- Han JD, Dupuy D, Bertin N, Cusick ME, Vidal M. (2005) **Effect of sampling on topology predictions of protein-protein interaction networks.** Nature Biotechnology, 23, 839-844.

- Hartwell H, Hopfield J, Leibler S and Murray AW. (1999) **From molecular to modular biology.** Nature 402, C47-C52.

- He X and Zhang J. (2006) **Why do hubs tend to be essential in protein networks?** Plos Genetics, 2, 6, e88.

- 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 SG, 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, Apweiler R. (2004) **The HUPO PSI's Molecular Interaction format-a community standard for the representation of protein interaction data.** Nature Biotechnology, 22, 2, 177-183

- Hernández P, Huerta-Cepas J, Montaner D, Al-Shahrour F, Valls J, Gómez L, Capellá G, Dopazo J, Pujana MA. (2007) **Evidence for system-level molecular mechanisms of tumorgenesis.** BMC Genomics, 8:186

- Hernandez-Toro J, Prieto C and De Las Rivas J. (2007) **APID2NET: unified interactome graphic analyzer.** Bioinformatics 23(18): 2495-2497

- Herrero J, Diaz-Uriarte R and Dopazo J. (2003) **An approach to inferring transcriptional regulation among genes from large-scale expression data.** Comparative and Functional Genomics, 4:148-154.

- Herrero J, Al-Shahrour F, Díaz-Uriarte R, Mateos A, Vaquerizas JM, Santoyo J, Dopazo J. (2003) **GEPAS: A web-based resource for microarray gene expression data analysis.** Nucleic Acids Res, 31, 3461-7.

- Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sørensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M. (2002) **Systematic identifiaction of protein complexes in Saccharomyces cerevisiae by mass spectrometry.** Nature, 415, 180-183

- Hu Z, Ng DM, Yamada T, Chen C, Kawashima S, Mellor J, Linghu B, Kanehisa M, Stuart JM, DeLisi C. (2007) **VisANT 3.0: new modules for pathway visualization, editing, prediction and construction.** Nucleic Acids Research, 1-8

- Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S,

Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E. (2007) **Ensembl 2007.** Nucleic Acids Res. Vol. 35, Database issue:D610-D617.

- Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L. (2001) **Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network.** Science, 292, 929-933

- Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. (2001) **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** Proc. Natl. Acad. Sci. 98, 4569-4574.

- Jansen R, Greenbaum D, Gerstein M. (2002) **Relating whole-genome expression data with protein-protein interactions.** Genome Research, 12, 37-46

- Johsson PF and Bates PA. (2006) **Global topological features of cancer proteins in the human interactome.** Bioinformatics, 22(18), 2291-2297.

- Joy MP, Brock A Ingber DE and Huang S. (2005) **High-Betweenness Porteins in the Yeast Protein Interaction Network.** Journal of Biomedicine and Biotechnology, 2, 96-103

- Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M. (2004) **The KEGG resource for deciphering the genome.** Nucleic Acids Res, 32, D277-280.

- Kelley R and Ideker T. (2005) **Systematic interpretation of genetic interactions using protein networks.** Nature Biotechnology, 23, 5, 561-566

- Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. (2006) **IntAct – Open Source Resource for Molecular Interaction Data.** Nucleic Acids Res.;35(Database issue):D561-5.

- Khatri P, Bhavsar P, Bawa G, Draghici S. (2004) **Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments.** Nucleic Acids Res.;32(Web Server issue):W449-56.

- Khatri P and Draghici S. (2005) **Ontological analysis of gene expression data: current tools, limitations, and open problems.** Bioinformatics, 21(18):3587-3595.

- Khatri P, Sellamuthu S, Malhotra P, Amin K, Done A, Draghici S. (2005) **Recent additions and improvements to the Onto-Tools.** Nucleic Acids Res.;33(Web Server issue):W762-5.

- Khatri P, Desai V, Tarca AL, Sellamuthu S, Wildman DE, Romero R, Draghici S. (2006) **New Onto-Tools: Promoter-Express, nsSNPCounter and Onto-Translate.** Nucleic Acids Res.;34(Web Server issue):W626-31.

- Khatri P, Voichita C, Kattan K, Ansari N, Khatri A, Georgescu C, Tarca AL, Draghici S. (2007) **Onto-Tools: new additions and improvements in 2006.** Nucleic Acids Res.;35(Web Server issue):W206-11.

- Kim SY and Volsky DJ. (2005) **PAGE: parametric analysis of gene set enrichment**. BMC Bioinformatics, 6:144.

- Kitano H. (2002) **Systems Biology: A Brief Overview.** Science, 295, 1662-1664

- Lee I, Lehner B, Crombie C, Wong W, Fraser AG, Marcotte EM. (2007) **A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans.** Nature Genetics, 40, 2, 181-188

- Lee RC, Feinbaum RL, Ambros V. (1993) **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** Cell 75: 843–854.

- Lee HK, Hsu AK, Sajdak J, Qin J and Pavlidis P. (2004) **Coexpression analysis of human genes across many microarray data sets.** Genome Res, 14, 1085-1094.

- Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M. (2004) **A map of the interactome network of the metazoan C. elegans.** Science, 303, 540-543

- Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, Kohane IS, Kasif S. (2007) **Network-Based Analysis of Affected Biological Proceses in Type 2 Diabetes Models.** Plos Genetics, 3, 6, e96.

- Liu X and Muller HG. (2003) **Modes and clustering for time-warped gene expression prole data**. Bioinformatics, 19(15):1937-1944.

- Luo F, Yang Y, Chen CF, Chang R, Zhou J, Scheuermann RH. (2007) **Modular organization of protein interaction networks.** Bioinformatics, 23, 2, 207-214

- Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA and Gerstein M. (2004) **Genomic analysis of regulatory network dynamics reveals large topological changes.** Nature , 431(7006):308-312.

- Maddox AM and Haddox MK. (1988) **Characteristics of cyclic AMP enhancement of retinoic acid induction of increased transglutaminase activity in HL60 cells.** Exp Cell Biol, 56, 49-59.

- Maere S, Heymans K, Kuiper M. (2005) **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.** Bioinformatics, 21(16), 3448-9

- Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. (2004) **GOToolBox: functional analysis of gene datasets based on Gene Ontology.** Genome Biol.;5(12):R101.

- Masseroli M, Martucci D, Pinciroli F. (2004) **GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining.** Nucleic Acids Res.;32(Web Server issue):W293-300.

- Masseroli M, Galati O, Pinciroli F. (2005) **GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists.** Nucleic Acids Res.;33(Web Server issue):W717-23.

- Mateos A, Dopazo J, Jansen R, Tu Y, Gerstein M, Stolovitzky G. (2002) **Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons.** Genome Res., 12(11):1703-1715.

- Mathivanan S, Periaswamy B, Gandhi TK, Kandasamy K, Suresh S, Mohmood R, Ramachandra YL, Pandey A. (2006) **An evaluation of human protein-protein interaction data in the public domain.** BMC Bioinformatics, 7, 5, S19

- Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. (2002) **Network motifs: simple building blocks of complex networks.** Science, 298, 824–827.

- Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U. (2004) **Superfamilies of evolved and designed networks.** Science, 303, 1538–1542.

- Mlecnik B, Scheideler M, Hackl H, Hartler J, Sanchez-Cabo F, Trajanoski Z. (2005) **PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways.** Nucleic Acids Res. ;33(Web Server issue):W633-7.

- Montaner D, Tárraga J, Huerta-Cepas J, Burguet J, Vaquerizas JM, Conde L, Minguez P, Vera J, Mukherjee S, Valls J, Pujana MA, Alloza E, Herrero

J, Al-Shahrour F, Dopazo J. (2006) **Next station in microarray data analysis: GEPAS.** Nucleic Acids Res, 34, W486-91.

- Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Merirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, and Groop LC. (2003) **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** Nat Genet, 34(3):267-273.

- Mrowka R, Patzak A & Herzel H. (2001) **Is there a bias in proteome research?** Genome Research, 11, 1971-1973

- Nau GJ, Richmond JF, Schlesinger A, Jennings EG, Lander ES and Young RA. (2002) **Human macrophage activation programs induced by bacterial pathogens.** Proc. Natl. Acad. Sci. USA, 99(3):1503-1508.

- Newman JC and Weiner AM. (2005) **L2L: a simple tool for discovering the hidden significance in microarray expression data.** Genome Biology 6:R81

- Pereira-Leal JB, Enright AJ, Ouzounis CA. (2004) **Detection of functional modules from protein-protein interaction networks.** Proteins, 54, 49-57.

- Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. (2003) **Development of human protein reference database as an initial platform for approaching systems biology in humans.** Genome Research, 13:2363-2371.

- Prieto C and De Las Rivas J. (2006) **APID: Agile Protein Interaction DataAnalyzer.** Nucl. Acids Res. 34: W298-W302

- Przulj N. (2006) **Biological network comparison using graphlet degree distribution**. Bioinformatics Vol 23 ECCB 2006, e177-e183.

- Ramoni MF, Sebastiani P and Kohane IS. (2002) **Cluster analysis of gene expression dynamics.** Proc. Natl. Acad. Sci. USA, 99(14):9121-9126.

- Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, Stark C, Ho Y, Botstein D,

Andrews B, Boone C, Troyanskya OG, Ideker T, Dolinski K, Batada NN, Tyers M. (2006) **Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae.** J. Biol 5:11.

- Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, Varambally S, Ghosh D, Chinnaiyan AM. (2007) **Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles.** Neoplasia; 9(2):166-80.

- Rives AW and Galitski T. (2003) **Modular organization of cellular networks.** Proc. Natl. Acad. Sci., 100, 3, 1128-1133.

- Robinson MD, Grigull J, Mohammad N, Hughes TR. (2002) **FunSpec: a web-based cluster interpreter for yeast.** BMC Bioinformatics.;3:35.

- Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. (2005) **Towards a proteome-scale map of the human protein-protein interaction network.** Nature 20;437(7062):1173-8.

- Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. (2004) **The Database of Interacting Proteins: 2004 update.** NAR 32 Database issue D449-51.

- Schena M, Shalon D, Davis RW, Brown PO (1995) **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** Science 270: 467–470.

- Schliep A, Schonhuth A, and Steinhoff C. (2003) **Using hidden Markov models to analyze gene expression time course data.** Bioinformatics, 19 Suppl 1:i255-i263.

- Schuler GD, Epstein JA, Ohkawa H, Kans JA. (1996) **Entrez: molecular biology database and retrieval system.** Methods Enzymol, 266, 141-162.

- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. (2003) **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** Genome Research, 13(11), 2498-504.

- Shen -Orr, Milo R, Mangan S, Alon U. (2002) **Network motifs in the transcriptional regulation network of Escherichia coli.** Nat Genet. 31(1):64-8.

- Smid M and Dorssers LC. (2004) **GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms.** Bioinformatics, 20, 2618-2625.

- Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D and Futcher B. (1998) **Comprehensive identication of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** Mol. Biol. Cell, 9(12):3273-3297.

- Sporn O, Honey CJ, Kötter R. (2007) **Identification and classification of hubs in brain networks.** Plos One 10, e1049.

- Stegmaier K, Ross KN, Colavito SA, O'Malley S, Stockwell BR, Golub TR. (2004) **Gene expression-based high-throughput screening(GE-HTS) and application to leukemia differentiation.** Nat Genet, 36, 257-63.

- Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. (2005) **A human protein-protein interaction network: a resource for annotating the proteome.** Cell 2005 122(6):957-68.

- Stuart JM, Segal E, Koller D, Kim SK. (2003) **A gene-coexpression network for global discovery of conserved genetic modules.** Science 302(5643):249-255.

- Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. (2008) **Estimating the size of the human interactome.** Proc. Natl. Acad. Sci., 105, 19, 6959-6964

- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. (2005) **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** Proc Natl Acad Sci U S A, 102, 15545-15550.

- Suderman M and Hallet M. (2007) Tools for Visually Exploring Biological Networks. Bioinformatics, 23(20):2651-9

- Tárraga J, Medina I, Carbonell J, Huerta-Cepas J, Mínguez P, Alloza E, Al-Shahrour F, Vegas-Azcarate S, Gotz S, Escobar P, García-García F, Conesa A, Montaner D and Dopazo J. (2008) **GEPAS, a web-based tool for micrarray data analysis and interpretation Nucleic Acids Res.** 2008 Jul 1;36(Web Server issue):W308-14.

- Tomfohr J, Lu J, Kepler TB. (2005) **Pathway level analysis of gene expression using singular value decomposition.** BMC Bioinformatics.;12;6:225.

- Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. (2005) **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** Nature, 403, 603-627

- Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. (1995) **Serial Analysis of Gene Expression.** Science 270 (5235): 484-7.

- Volinia S, Evangelisti R, Francioso F, Arcelli D, Carella M, Gasparini P. (2004) **GOAL: automated Gene Ontology analysis of expression profiles.** Nucleic Acids Res.;32(Web Server issue):W492-9.

- von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. (2002) **Comparative assessment of large-scale data sets of protein-protein interactions.** Nature, 417, 399-403.

- von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, Bork P. (2007) **STRING 7 – recent developments in the integration and prediction of protein interactions.** Nucleic Acids Research, 35, Database issue, D358-D362.

- Wachi S, Yoneda K and Wu R. (2005) **Interactome-trancriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues.** Bioinformatics, 21(23), 4205-4208.

- Westfall, P H and Young, S S. (1993) **Resampling-based multiple testing.** John Wiley & Sons. New York.

- Wilkinson DM and Huberman BA. (2004) **A method for finding communities of related genes.** Proc. Natl. Acad. Sci., 101, 1, 5241-5248.

- Wolfe CJ, Kohane IS and Butte AJ. (2005) **Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks.** BMC Bioinformatics, 6:227.

- Wu G, Fang YZ, Yang S, Lupton JR, Turner ND. (2004) **Glutathione metabolism and its implications for health.** J Nutr, 134, 489-92.

- Xu Q, Simpson SE, Scialla TJ, Bagg A, Carroll M. (2003) **Survival of acute myeloid leukemia cells requires PI3 kinase activation.** Blood, 102, 972-80.

- Xenarios I and Eisenberg D. (2001) Protein interaction databases. Current Opinion in Biotechnology, 12:334-339

- Xu J and Li Y. (2006) **Discovering disease-genes by topological features in human protein-protein interaction network.** Bioinformatics, 22, 2800-2805.

- Yeger-Loten, E, Sattah, S Kashtan, N, Itzkovitz, S, Milo, R, Pinter, R, Alon, U Margalit, H. (2004) **Networks motifs in integrated cellular networks of transcription-regulation and protein-protein interactions.** PNAS 101, 16, 5934-5939.

- Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN. (2003) **GoMiner: a resource for biological interpretation of genomic and proteomic data.** Genome Biol, 4, R28.

- Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN. (2005) **High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID).** BMC Bioinformatics;6:168.

- Zhang B, Schmoyer D, Kirov S, Snoddy J. (2004) **GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies.** BMC Bioinformatics;5:16.

- Zhang B, Kirov S, Snoddy J. (2005) **WebGestalt: an integrated system for exploring gene sets in various biological contexts.** Nucleic Acids Res.;33(Web Server issue):W741-8.

- Zhu G, Spellman PT, Volpe T, Brown PO, Botstein D, Davis TN, and Futcher B. (2000) **Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth.** Nature, 406(6791):90-94.

"Will nature make a man of me yet?"
This Charming Man, Morrissey, 1983.