Facultad de Matemáticas
Departamento de Estadística e Investigación Operativa

UNDERSTANDING DISEASE MECHANISMS WITH STATISTICAL MODELS OF SIGNALING PATHWAY ACTIVITIES

Thesis submitted by
Patricia Sebastián León

Supervised by
Joaquín Dopazo

Tutor
Carmen Armero

2016

Facultad de Matemáticas
Departamento de Estadística e Investigación Operativa

UNDERSTANDING DISEASE MECHANISMS WITH STATISTICAL MODELS OF SIGNALING PATHWAY ACTIVITIES

Patricia Sebastián León

Director:
Dr. Joaquín Dopazo

Tutora:
Dra. Carmen Armero

2016

DECLARACIÓN AUTORÍA

JOAQUÍN DOPAZO BLÁZQUEZ, doctor en Biología por la Universidad de Valencia y jefe del departamento de Genómica Computacional del Centro de Investigaciones Príncipe Felipe (CIPF), director del grupo de Bioinformática en Centro de Investigación Biomédica en Red Enfermedades Raras (CIBERER) y director del nodo de Genómica Funcional en el Instituto Nacional de Bioinformática (INB)

CERTIFICO

Que la presente memoria con el título **"UNDERSTANDING DISEASE MECHANISMS WITH STATISTICAL MODELS OF SIGNALING PATHWAY ACTIVITIES"** ha sido realizada por Dña. **PATRICIA SEBASTIÁN LEÓN** bajo mi dirección y reúne todos los requisitos para su presentación y defensa como TESIS DOCTORAL ante un tribunal.

Y para que así conste a los efectos oportunos, firmo la presente certificación en Valencia a 15 de Diciembre de 2015.

Fdo. Dr. Joaquín Dopazo Blázquez

Agradecimientos

Nunca pensé que llegaría este momento, escribir la última parte de esta tesis, los agradecimientos. Me ha costado mucho decidir como hacerlo. Son muchas las personas que han puesto su granito de arena para que este barco llegara a buen puerto, y esta tesis tiene una parte de cada uno de ellos, los que están y los que, por una razón u otra, ya se han ido. No os puedo incluir uno a uno en estos agradecimientos, tendría que escribir otra tesis :) pero espero haberos hecho saber a cada uno de vosotros lo que ha significado vuestra ayuda para cumplir este sueño, que esta tesis viera la luz. He tenido muchos momentos en los que no me veía capaz, muchos momentos en los que hubiera dejado todo y me hubiera escapado a una isla desierta, pero nunca me habéis dejado, y nunca, nunca, podré agradeceroslo lo suficiente.

Gracias Ximo por el apoyo personal y profesional, gracias por darme la oportunidad de aprender y crecer junto a ti. Gracias Ana por abrirme las puertas del maravilloso y emocionante mundo de la bioinformática.

Gracias a toda mi familia por el cariño y la comprensión. Gracias Tete y Pili por tener siempre un abrazo listo para mi, Gracias Christian y Claudia, mis encantadores sobrinos, por llenar de alegría mi mundo. Gracias a mis padres por estar siempre ahí. Por apoyarme y animarme cuando decidí empezar este viaje. Por quererme incondicionalmente.

Gracias a mis compis bioinfos. Por enseñarme todo lo que se y por todas las horas compartidas, dentro y fuera del labo. Sois un grupo estupendo y espero que nuestro camino juntos no acabe aquí.

Y, sobretodo, gracias a mis amigos, a todos. Gracias por levantarme cuando creía que no podía más. Gracias por enseñarme el lado bueno de las cosas cuando yo sólo podía ver negro. Gracias por reñirme cuando no tenía razón. Gracias por los sabios consejos. Gracias por hacerme reír siempre. Gracias por no dejarme sentir sola nunca. Gracias por los momentos compartidos por el mundo. Gracias. Sabeis quienes sois, sin vosotros esta tesis no hubiera sido posible.

Abstract

Understanding the processes that cause diseases is now one of the hot topics in biological research. In the past, most of the genetic diseases were associated to a single gene, but there are a lot of diseases that cannot be explained by the action of a single gene, and they can only be explained, but by malfunctions of a set of genes. After human genome was sequenced, it became evident that genes do not act alone in a cell, but together in an intricate network of relationships that determine their activity. This discover resulted in the beginning of systems biology, that aims to understand how cellular components are related to give rise to life. In systems biology, networks describing the relationships that can be established between gene are called signaling pathways.

The explosion of new high-throughput technologies has allowed to measure simultaneously thousands of cellular elements, including gene expression, so many approaches trying to explain the differential behavior of these signaling pathways when comparing control and disease samples by using gene expression data have appeared in the last years.

In this thesis, a novel methodology is introduced to analyze expression data in a pathway context. First, signaling pathways were modeled and dissected in smaller substructures called subpathways, that collect the individual biological functions included into the pathway. This model has into account the different nature of nodes and edges in the pathway network and evaluates them according to the underlying biology that they represent. The levels of expression of several probe sets from different microarray platforms were also modeled. Specifically, probe set distribution was fitted to a mixture of two distributions (gamma or normal) associated to the active and inactive state of the probe set. These distributions were then used to estimate the probability of a probe set to be active in a given sample. The estimated probability of activation was first summarized to node activation probability and then was propagated along the subpathways, by having into account the different types of edges that relate the nodes; this results in an estimation of the activity of each subpathway. Given

a control-case experiment, the proposed approach estimates the activity of each subpathway in each sample, and compare then to obtain the subpathways significantly activated/deactivated between both conditions.

This method overcomes most of the limitations of previous methods and provides the research community with a user-friendly web-tool to analyze expression data from a control-disease experiment in a signaling pathway context. It is very useful for an easy interpretation of the results in terms of gain/loss of biological functions. Thus, with this strategy, it is possible to understand the mechanisms driving a disease in terms of systems biology, since we obtain a comprehensive path of genes, closely related, that produce a determined function in the cell, instead of a single gene or a set of genes without relation between them.

Resumen

Hoy en día, uno de los temas más candentes en la investigación biomédica es entender los procesos que producen enfermedades. En el pasado, la mayor parte las enfermedades génicas fueron asociadas al mal funcionamiento de un solo gen, pero hay muchas enfermedades que sólo pueden ser explicadas por el mal funcionamiento de un conjunto de genes. Una vez el genoma humano fue secuenciado, se hizo evidente que los genes no actúan solos en la célula, si no que están unidos por una intrincada red de interacciones que determinan su actividad. Este descubrimiento dio lugar al inicio de la biología de sistemas, que trata de entender cómo los componentes celulares se relacionan entre ellos para dar lugar a la vida. En biología de sistemas, las redes que describen las relaciones que se establecen entre genes se llaman rutas de señalización. Por otra parte, el auge de nuevas tecnologías de alto rendimiento, ha permitido en los últimos años medir simultáneamente miles de componentes celulares, incluyendo la expresión génica. Esto ha producido la creación de nuevas metodologías para el estudio del comportamiento diferencial de las rutas de señalización entre dos condiciones experimentales dados los niveles de expresión de cada uno de los genes contenidos en la ruta. En esta tesis, presentamos una nueva metodología para analizar los datos de expresión en el contexto de las rutas de señalización. Primero, las rutas de señalización fueron modeladas y divididas en subestructuras más pequeñas, que llamaremos subrutas, que recogen las diferentes funciones biológicas individuales incluidas en la ruta completa. Este modelado tiene en cuenta la diferente naturaleza tanto de los nodos como de las aristas que forman la red, permitiendo que sean evaluadas de acuerdo con el concepto biológico que representan. Además, se modelaron las distribuciones de niveles de expresión de las sondas de diversos chips the expresión génica. Concretamente, la distribución de cada sonda fue ajustada a una mixtura de distribuciones (gamma o normal) asociadas con los estados activo e inactivo de la sonda, respectivamente. Estas distribuciones fueron usadas para estimar la probabilidad de que una sonda esté activa en una determinada muestra. La

probabilidad de activación estimada fue propagada a continuación a lo largo de las subrutas, teniendo en cuenta las diferentes relaciones que se pueden establecer entre los nodos, obteniendo una estimación de la actividad de cada una de las subrutas. Por lo tanto, dado un experimento que compara dos condiciones biológicas, la metodología propuesta estima la actividad de cada una de las subrutas, y las compara para obtener las subrutas significativamente activas o inhibidas entre ambas condiciones. Esta metodología supera la mayor parte de las limitaciones presentadas por los métodos anteriores y proporciona a la comunidad científica una herramienta web de fácil manejo que permite analizar los datos de expresión obtenidos en un experimento comparando dos condiciones dentro del contexto de las rutas de señalización. En consecuencia, la estrategia propuesta en esta tesis nos permite entender los mecanismos que dan lugar a una enfermedad en términos de la biología de sistemas, ya que permite obtener como resultado de nuestro análisis un conjunto de genes, biológicamente relacionados y que conjuntamente producen una determinada función génica, en vez de un gene o conjunto de genes sin ninguna relación entre ellos.

Contents

Chapter

1 Introduction

"Begin at the beginning," the King said gravely, "and go on till you come to the end: then stop."

— Lewis Carroll, *Alice in Wonderland*

1.1 Molecular biology

Biology can be seen as a counterpoint between two different elements: individual variability and the constancy of fundamental mechanisms. The divergence of life created a huge variety of living forms, but the fundamental mechanisms were conserved in the evolutionary process. Therefore, studying these mechanisms essential to understanding the reasons for the differences and similarities between living things.

One of the lowest levels of organization in living organisms is the cell. Unicellular organisms are formed by one single cell, while multicellular organisms are comprise groups of cells performing specialized functions and are linked by intricate communication systems. Strikingly, all the information needed to generate the molecules that perform cellular functions, is stored in the form of double-stranded deoxyribonucleic acid (DNA) molecules. DNA is the prerequisite for the inception of all higher life forms for two main reasons: first, it allows species to pass traits from parents to their offspring, thus producing individuals with the same characteristics, and second, it controls the production of the different proteins that perform most of the function of the cells.

The structure of DNA was discovered by the researchers James Watson and Francis Crick en 1953[1], who jointly received the Nobel Prize in Physiology or Medicine in 1962 for their discovered. It consists of long unbranched, paired polymer chains, formed by four types of monomers, called nucleotides which are named depending on the base that forms them: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). The doubled-stranded structure of the DNA molecule is organized according base complementation: A binds to T, and C binds to G. The complete set of information carried by an organism in its complete DNA sequence is called the genome and it carries information about all the proteins that an organism will ever synthesize. A huge amount of information is contained in the genome is astonishing, for example, using the four-letter nucleotide alphabet, the nucleotide sequence for a small gene can take up a page of text, and the whole human genome would fill thousands of books.

The discovery of DNA, which was not originally recognized for the breakthrough it was, allows researches to understand the relationship between DNA and the final protein product that performs determined processes in the cell. This process was described by Crick in 1970 (Figure 1.1) in the *Central Dogma of Molecular Biology*[2]. This dogma describes a two-step process, known as gene expression. First, DNA is transcribed into ribonucleic acid (RNA) molecules and then, RNA is translated into proteins. This description is ex-

tremely simplistic, but serves as a very useful for understanding the flow of information in biological systems. The details of the process have been elaborated by later discoveries which introduce regulatory components and changes in the flows of information. Here only the modifications which are necessary for understanding this thesis will be introduced[1].

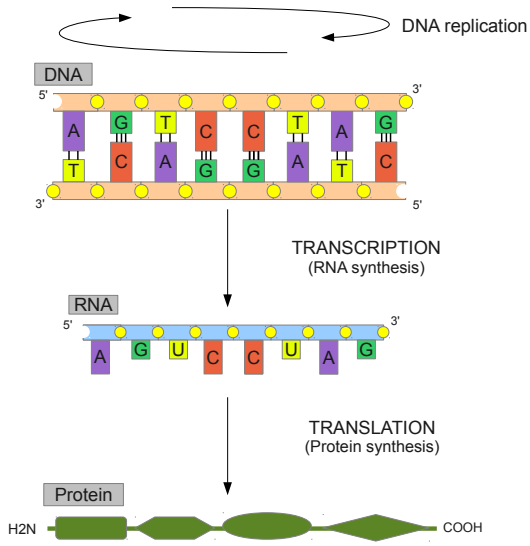


Figure 1.1: ***Central Dogma of Molecular Biology.*** Information encoded in deoxyribonucleic acid (DNA) is used to synthesize ribonucleic acid (RNA) molecules in a process called transcription. These RNA molecules then mediate protein synthesis in a process termed translation. The information contained in DNA molecules can also be used to generate other DNA molecules in the process of DNA replication

In the first step in the *Central Dogma of Molecular Biology*, called transcription, DNA is transcribed into RNA molecules. DNA molecules are usually very large and contain the specifications for thousands of proteins; it is not transcribed as a whole, rather, individual segments of the entire DNA sequence

[1] For more details see [3]

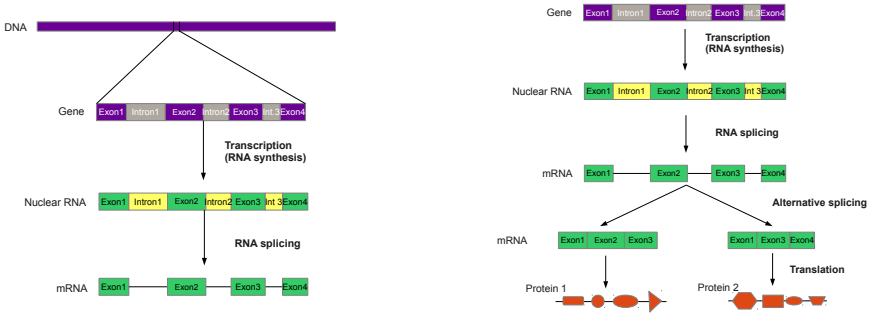
are transcribed into separate molecules, each one is translating into different proteins. Each DNA segment that encodes a protein is known as a gene: during transcription, genes are used as a template to synthesize shorter molecules RNA polymers. These molecules are very closely related to DNA, except for two characteristics: RNA is a single-stranded molecule, and Uracil (U) substitute the Thymine (T) base.

In second step, called translation, RNA molecules direct the synthesis of another type of polymers: proteins. Protein molecules, are also long, unbranched polymer chains formed by monomers. However, protein monomers are different from those comprising DNA and RNA. In the translation process, information encoded in the RNA sequence is read out in groups of three nucleotides at a time, each nucleotides triplet (codon) specifies a single amino acid in the resulting protein. There are 64 (=4x4x4) possible triplet combinations, and all of them occur in nature, but several codons correspond to the same amino acid (see Table 1.1); there are 20 types of amino acids.

1st	2nd position U		2nd position C		2nd position A		2nd position G		3rd
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG		UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Table 1.1: **Nucleotide triplet to aminoacid conversion table.** Depending on the position of the nucleotides in the triplet, different aminoacids are synthesized. (The abbreviations of the aminoacid name are shown in the abbreviations section).

One of the most relevant modifications of the *Central Dogma of Molecular Biology* came with the discovery of RNA splicing. In 1977, Sambrook discovered that genes are composed of two types of regions called introns and exons[4]. Exons are regions of DNA that code for a protein, while introns do not code for any specific protein and are spliced out of RNA molecules before they are translated, giving rise to a final messenger RNA (mRNA), which only contain information from coding exonic regions. This process is referred as RNA splicing (Figure 1.2a). In 1978, Gilbert described the alternative splicing process of transcription in which introns are dismissed and exons are combined in different ways in order to produce different proteins. That is, one gene can synthesize several different proteins by combining different exon, and each combination produces a protein called a transcript. Therefore, a gene can be translated into several transcripts that encode different proteins (Figure 1.2b).



(a) RNA splicing

(b) Alternative splicing

Figure 1.2: **(a) RNA splicing.** A gene is formed by exons (coding regions of DNA) and introns (non-coding DNA regions) . Each gene is transcribed in its entirety into nuclear RNA and then the RNA splicing process discards the intron regions to form mRNA comprising only coding regions. **(b) Alternative splicing.** After RNA splicing, alternative splicing combines different exons to produce different transcripts that leads to the synthesis of different proteins.

1.2 Gene expression

Gene expression is the primary mechanism by which the genotype (hereditary information contained in the DNA) gives rise to the phenotype (genotype expression modified by the environment). Gene expression is an interpretation of the genetic code, in the sense that the properties of the products of gene expression lead to the organism's phenotype.

Gene expression can be measured at different levels (usually at the mRNA or protein levels), corresponding with both steps of the *Central Dogma of Molecular Biology* (Figure 1.1). It is currently cheaper, and more precise and comprehensive to measure mRNA levels and this have the added benefit of compatibility with the use of high-throughput techniques for measuring thousands of genes in a single run. Even though mRNA is sometimes not the ultimate product of a gene, and correlation between mRNA and protein levels is not always straightforward[5, 6], most approaches use mRNA measurements as reasonably accurate proxies for protein behavior[7]. Specifically, most of them use differential expression to compare the abundance of mRNA produced by cells in two or more different conditions.

The way in which mRNA levels are measured has evolved since its beginnings in the 70s when the northern blot technique was first developed by Alwine and colleges[8]. Northern blotting is a technique for measuring gene expression by detecting RNA (or isolated mRNA) in a sample[8]. This technique allows gene expression to be measured in different conditions, tissues, or disease states and compared with a control, but it only allows one or a small number of genes to be measured at once.

In the 80s, Kary Mullis developed a revolutionary method called PCR (polymerase chain reaction)[9] which superceded northen blotting and for which he received the Nobel Prize for Chemistry in 1994. This technique allowed researches to amplify any selected piece of DNA by several orders of magnitude, to generate thousands to millions of copies of a particular DNA sequence. Based on this technique, several methods for measuring gene expression were developed including reverse transcription PCR (RT-PCR), quantitative PCR, real time PCR and real time qRT-PCR (real time quantitative reverse transcription PCR). These techniques are highly quantitative and sensitive, and are generally used for interrogating a small number of genes in a large number of samples.

Later, in the 90s, high density oligonucleotide microarray technology was developed to allow the simultaneous measurement of thousands of genes. These microarrays are formed by thousand of spots, each containing short DNA sequences known as probes, which are complementary to the target sequence of

interest. This technology is based on the capacity of nucleotides to specifically pair with their complementary base. Target probes (in the sample) are fluorescence labeled with, and then hybridized with probes in the microarray. The higher the number of target probes the more intense the fluorescence signal is, which consequently, serves as a proxy for the expression level of the probe. Several companies have developed microarrays for monitoring known genes from a genomic point of view.

More recently, the increased use and decreased price of high-throughput sequencing technologies has allowed steady-state RNA in samples to be sequenced. This technique, known as RNA-seq, is "unbiased" (free of dependency on prior knowledge of the sequence of the organism), making it ideal for discovery, but it remains costly and difficult computationally speaking. Several comparisons of RNA-seq and microarray measurements have been performed, which have shown that there is strong agreement between both data sets [10]. Even though RNA-seq is more sensitive, microarrays continue to be far more successful in many biological settings[11].

1.2.1 Affymetrix DNA microarray gene expression data

DNA microarray gene expression technology is widely used and there are several companies developing different chips for monitoring every known gene from a genomic point of view. Affymetrix GeneChip arrays[12] are the most popular and they are used by thousands of researchers around the world[2].

A DNA microarray is a collection of DNA spots attached to a solid surface. Each DNA spot contains picomoles of specific short DNA sequences known as probes. In the case of Affymetrix technology these probes are oligonucleotides (short sequences of nucleotides) which are 25 bp long. Any mRNA molecule of interest (related to a gene) is represented by several probe sets, each of them comprising 16-20 probe pairs. Each probe pair is composed of a perfect match (PM) probe and a mismatch (MM) probe. The PM probe represents a section of the mRNA molecule of interest and the MM probe is created by changing the middle (13th) nucleotide of the PM probe sequence[3] (Figure 1.3).

Each microarray is formed by millions of spots, each of them containing many copies of a specific DNA sequence (Figure 1.4b). To measure gene expression in a sample, the RNA in the cell sample is extracted and all the target probes in the sample are marked with a fluorescence dye and hybridized onto

[2] A simple search in PubMed `www.ncbi.nlm.nih.gov/pubmed` returns more than 7,000 papers citing this technology

[3] This is done to measure non-specific binding

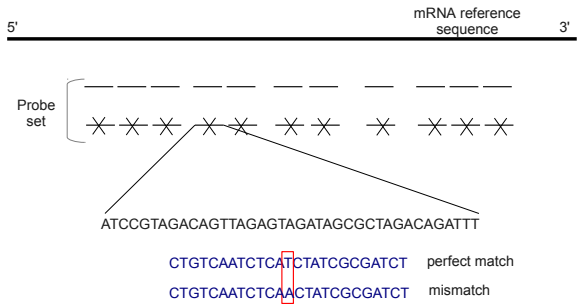


Figure 1.3: **Affymetrix probe set** Several probes are selected from a mRNA reference sequence to represent the gene. These probes are paired: there is a perfect match sequence as well as a mismatch sequence in which only the central nucleotide differs.

the array (Figure 1.4c). After scanning the array, intensity values PM_{ij} and MM_{ij} are recorded for arrays $(i = 1, ..., I)$ and probe pairs $(j = 1, ..., J)$ for any given probe sets in the array. The highest intensity value represents the most expressed probe sequence in the sample (Figure 1.4d).

The intensities of the spots represents the degree of hybridization for each oligonucleotide probe. However, systematic biases are included in expression level measurements for several reasons, for instance the inclusion of unequal quantities of RNA or differences in the labeling or detection efficiencies between the fluorescent dyes[13]. Therefore, the fluorescence intensity of each spot must be normalized by adjusting individual hybridization intensities before meaningful biological comparisons. The intensities of the 16-20 probe pair sets also have to be summarized into a single value in order to define a single expression measurement that represents the amounts of mRNA for each probe set; there exist many methods to do this, the most common of which are described below.

The intensities of each probe pair are defined as[14]:

$$PM_{ijn} \text{ and } MM_{ijn},\ i = 1, ..., I; j = 1, ..., J; n = 1, ..., N$$

where: n represents the different genes, i the RNA samples and j the probe pair number.

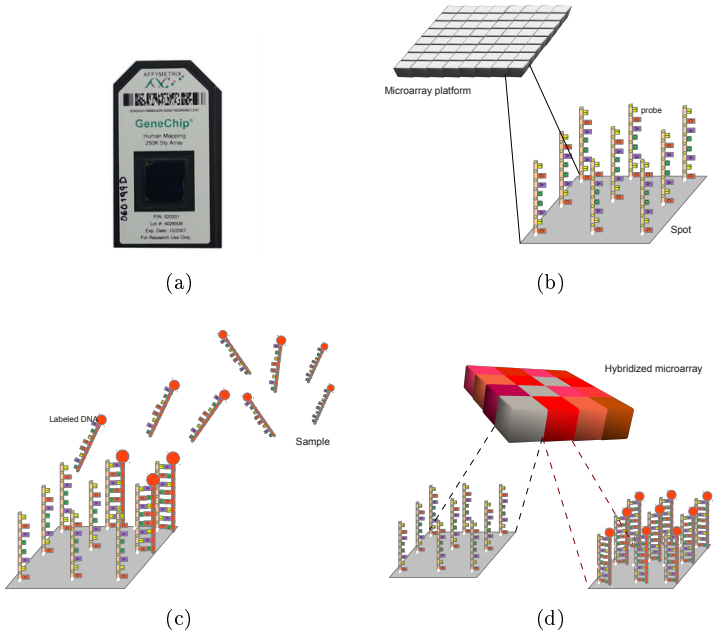


Figure 1.4: **Affymetrix microarray functioning (a)** An Affymetrix microarray. **(b)** A microarray can contain thousands of spots, each of them contains many copies of a particular DNA sequence corresponding to a probe. **(c)** An RNA sample is extracted from the cell, labeled with a fluorescent dye, and hybridized with the complementary sequence. **(d)** Dyes are illuminated using fluorescent light to show which RNA fragments hybridized to in which microarray spots, and consequently which probes are active in the cell.

The four commonly-used summary density measurements are:

Affymetrix's average difference (`AvDiff`) This is the Affymetrix default. It is a trimmed mean defined as:

$$AvDiff = \frac{1}{\#A} \sum_{j \in A} (PM_j - MM_j)$$

where A the subset of probes for which $d_j = PM_j - MM_j$ are within three

standard deviations of the average of $d_{(2)}, .., d_{(J-1)}$. This measurement has some limitations: $MM > PM$ in about one third of the probes and the linear scale measurement is not optimal for this type of data.

Multiplicative model-based expression index (dChip) The dChip[15] is defined as the maximum likelihood estimates at the θ_i, $i = 1, ..., I$ obtained from fitting the following model:

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \epsilon_{ij}$$

where: ϕ_j represents probe-specific affinities. Also, the noise component is assumed to have independent normally distributed errors ($\epsilon_{ij} \sim N(0, \sigma^2)$).

MAS 5.0 signal This robust average measurement was developed by Hubbell[16] and is defined as:

$$signal = TukeyBiweight(log(PM_j - CT_j))$$

where CT_j is a "repaired" version of MM defined as:

$$CT_j = \begin{cases} MM_j & \text{if } MM_j < PM_j \\ \text{less than } PM_j & \text{if } MM_j \geq PM_j \end{cases}$$

Robust Multi-array Average (RMA) This methodology, developed by Izarry[14], describes a process in which raw intensity values are background corrected, log2 transformed and quantile normalized. A linear model is then used to fit normalized data and to obtain an expression measurement for each probe set on each array. This method assumes that n, the background-adjusted, log2 transformed and normalized PM intensities (Y) follow a linear additive model for each probe set:

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \epsilon_{ijn}, \ i = 1, ..., I; j = 1, ..., J; n = 1, ..., N$$

where μ_{in} represents the log scale expression level for array i (and therefore its estimates give the expression measurements for probe set n on array j), α_{jn} represents the probe affinity effect[4] and ϵ_{ijn} represents an independent identically distributed error term with a mean of zero. A robust linear fitting procedure was used to estimate μ_{in} and the resulting summary statistic is referred to as RMA. It presents some advantages regarding the other defined measurements[14]:

[4] It is assumed that $\sum_j \alpha_j = 0$ as Affymetrix technology has chosen probes with intensities that on average are representative of the associated gene expression

1. Better precision, in particular for lower expression values.
2. It provides more consistent estimates of fold-change.
3. Higher specificity and sensitivity when using fold-change analysis.

1.2.2 Gene status from gene expression

Identifying which state a gene or probe set is in is a problem broadly studied in biology. Classification of genes in alternative states is a simplification of more complex patterns of gene behavior and action. Nevertheless, empirical evaluation of the observed data finds that gene expression patterns commonly fit one of two alternative expression level distribution states: present/absent (up/down). However, this classification is not easy since the range of expression of each gene and its activation/deactivation threshold can be extremely variable (Figure 1.5).

A expression array experiment result is a *Present/Absent call*[17] related to the MAS 5.0 measurement. This is based on a hypothesis test using PM and MM intensities. It supposes that the MM intensities are accurate estimates of a genes' specific background and it uses Wilcoxon signed-rank test**??** to declare if a gene is present or not. The decision rule is based upon the resulting p-values:

$$\begin{cases} \text{if } p < \alpha_1 & \text{Present (P)} \\ \text{if } \alpha_1 < p < \alpha_2 & \text{Marginally present (M)} \\ \text{if } p > \alpha_2 & \text{Absent (A)} \end{cases}$$

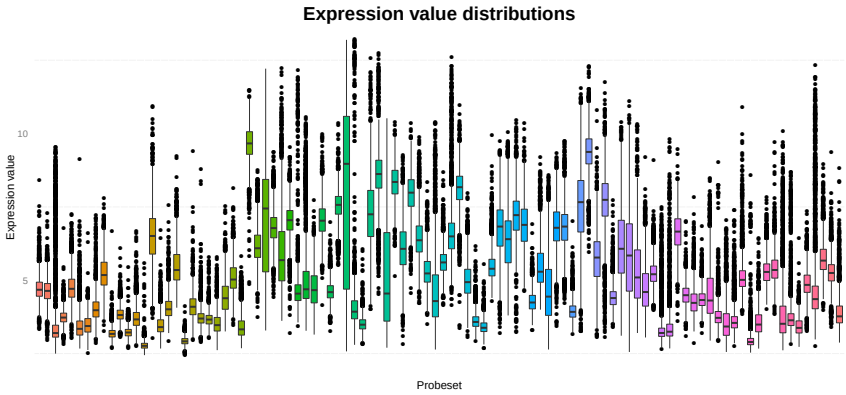
This approach, is often used to filter false positives from a large collection of probes on an expression array. But it has three principal disadvantages[18]:

1. It does not provide the user with a statistical gauge of the basic claim behind the *Present/Absent* call

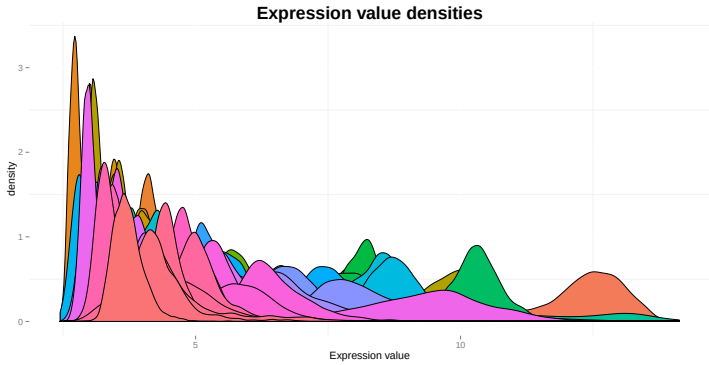
2. It does not compare calls across multiple samples

3. It cannot benefit from the increasingly sophisticated techniques for adjusting gene expression reading because it cannot operate on adjustable readings.

These limitations were overcome by Efroni[7] who developed a method to calculate the probability of each gene belonging to each state, a methodology which the same authors[18] better explained by using quantitative expression level readings to build a statistical expression model for each probe set, based



(a) Boxplots for 100 random probe sets



(b) Densitiy plots for 100 random probe sets

Figure 1.5: **Distribution of the expression of 100 random probe sets**. A different range of expression for each probe set can be observed. Short and long ranges of expression and high and low levels of expression can be seen, depending on the probe set.

on an assumed bimodal gene expression distribution that can be captured by mixed gamma distributions. This bimodal distribution is chosen after applying the Bayesian Information Criterion (BIC)[19]. The best fit is obtained with two distributions that account for the two states of an expressed gene: *Up*

and *Down*[7]. Therefore, data points can be used, for each single probe set, to infer a mixture of two gamma distributions, one distribution representing the *Down* state and another one representing the *Up* state. Once, these models are built for each probe set on the array and are given a new expression value, the probability of the new value belonging to each distribution can be estimated.

1.3 Bioinformatics and systems biology

Over the past few decades, advances in genomic technologies have produced explosive growth in the generation of biological data, which has also led to an unprecedented requirement for databases to store, organize, analyze and index the data, besides the fact that specialized tools are required to view and organize the data. This has resulted in new collaborations between computer scientists, statisticians, mathematicians, and engineers in order to process biological data, leading to the emergence of the field of bioinformatics, whose principal aims are[20]:

- To organize data in a way that allows researchers to access existing information and submit new entries.
- To develop tools and resources to analyze this of biological data.
- To use the tools developed to analyze data and interpret the results in a biologically meaningful manner.

Bioinformatics applications can be classified into three groups: sequence analysis, functional analysis, and structural analysis (Table 1.2). The creation of genomic projects focused on describing gene and protein functions and interactions caused the emergence of functional analysis as major field in the application of Bioinformatics. In this field there are a huge number of expression analysis datasets available in several conditions to the research community for several conditions, making it easier to discern how gene expression patterns change as a consequence of physiological cues and disease. The application of this knowledge applied to fields such as medicine or pharmacology allows both the causes of diseases and the development of drugs to treat them to be better understood. However, it will only be possible to analyze the gene expression profile of specific patients with a view to developing personalized drug for specifically for them when the price of this technology becomes more affordable.

Analysis	Description	Applications
Sequence	Analytical methods to understand the features, function, structure, or evolution of proteins, RNA and DNA.	Genome Annotation Comparative Genomics Genetics of Disease Oncogenomics Phylogeny Gene & Promoter Prediction
Functional	Analysis of the function and prediction of the functional interaction between various proteins or genes.	Gene expression studies Analysis of regulation processes Pathway modeling Protein interaction prediction Protein profiling Systems biology
Structure	Analysis and prediction of the three-dimensional structure and roles of macromolecules such proteins, RNA, and DNA.	Nucleic acid structure prediction Protein structure prediction Protein structure classification Protein structure comparison

Table 1.2: **Subfields in bioinformatics**

An emergent topic of functional analysis is known as Systems biology, which focuses on the complex interactions between genes and proteins within biological systems. Once the complete human genome sequence became known by Collins[21] and Venter[22],and genes and proteins were first listed, it became evident that the complexity of the human organism cannot be solely explained by the number of genes. Thus, the scientist started to ask questions about the gene interactions that could allow all the different different complex cell functions to be performed. This gave rise to the identification of an intricate network of relationships that determine their activity[23], including nodes representing molecules (genes, proteins, or different types of RNA) or metabolites (e.g sugars or lipids) and edges representing interactions between them.

Different types of biological networks exit, depending on the types of nodes and edges described by them. Two main classes can be established based on the interactions that they represent: genetic networks, which account for genetic linkages among genes, and physical networks, which account for physical,

functional, or regulatory relationships. Centering on physical networks, we can distinguish three main types of networks depending on the data they represent:

Gene regulatory networks They represent transcriptional and post-transcriptional gene expression regulation as a collection of elements which interact with each other indirectly to govern mRNA and protein expression levels.

Protein-protein interaction networks They represent protein-protein interactions, that is, two or more proteins which bind together, often to carry out their biological function.

Pathway networks They represents a series of actions among molecules in a cell that leads to production of a certain product or a cellular change. A pathway can trigger the assembly of new molecules, such as a fat or protein. Biological pathways can be divided into two main groups: metabolic pathways and signaling pathways. Metabolic pathways are series of chemical reactions occurring within a cell. Its main components are enzymes and metabolites. In a metabolic pathway a metabolite is converted into another metabolite in a series of chemical reactions catalyzed by enzymes. Signaling pathways are an entire set of cell changes induced by receptor activation. They are composed by genes and compounds (metabolites). Cell signaling is part of a complex system of communication that governs basic cellular activities and coordinates cell actions. In a signaling pathway the signal is transmitted from a stimulus gene to a receptor gene by activating or deactivating intermediate genes. This receptor gene synthesizes a certain protein that performs a predetermined function in the cell.

1.4 Data Repositories

One of the aims of bioinformatics is organize data in a way that allows researchers to access existing information and to submit new entries. Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analyses [24]. There are different types of databases depending on the information they contain (Table 1.3), but they all all share the following characteristics:

- They handle and share large volumes of biological data.

- They support large-scale analysis efforts.
- They make data accessible and updated.
- They link the knowledge obtained from various fields.

Database	Type of information
Bibliographic	Literature
Taxonomic	Classification of species
Genomic	Gene level information
Protein	Protein level information
Public data	Experimental data and results
Pathways	Metabolic and signaling pathways
Regulation	Transcription factors and events
PPIs	Protein-protein interactions

Table 1.3: **Biological databases classification**. Classification of databases according of the type of information they contained.

The databases used in the development of this thesis are described below. The first one is a repository of public data, Gene Expression Omnibus Database (GEO), where data used for analyses were downloaded. The second one is a pathway database, Kyoto Encyclopedia of Genes and Genomes (KEGG), from where information about pathways were extracted.

1.4.1 Gene Expression Omnibus database

The Gene Expression Omnibus (GEO) database is a public functional-genomics data repository from the NCBI (National Center for Biotechnical Information) [25, 26] that archives and freely distributes microarray, next-generation sequencing and other forms of high-throughput functional genomic data submitted by the scientific community. Tools are provided to help users query and download experiments and curated gene expression profiles. The three main goals of GEO are to:

- Provide a robust, versatile database in which high-throughput functional genomic data can be efficiently stored.

- Offer simple submission procedures and formats that support complete and well-annotated data deposits from the research community.
- Provide user-friendly mechanisms that allow users to query, locate, review and download studies and gene expression profiles of interest.

GEO is now one of the most populars databases for sharing experimental data with the research community and stores more than one million samples which are available for free. GEO information is organized in various levels. Samples (individual data) are linked together by series, that relate them to provide a focal point for the whole study. Series are grouped into datasets, that represent a a curated collection of biologically and statistically comparable samples/series. Finally, platforms represent the different technologies which the data comes from (Table 1.4)

Entity	Number of entries
Datasets	3,848
Series	52,901
Platforms	16,661
Samples	1,291,142

Table 1.4: **Some numbers in the Gene Expression Omnibus (GEO) database.** This table shows the number of entries at the different levels of the database's organization (December, 2014)

1.4.2 Pathway databases

The Kyoto encyclopedia of Genes and Genomes (KEGG)[27, 28] is a database resource used to understand the high-level functions and utilities of biological systems, such as a cell, an organism, or an ecosystem, from genomic and molecular-level information. It integrates sixteen main databases in four categories (Table 1.5)

The information contained in the PATHWAY database is classified into seven groups: *Metabolism*, *Genetic Information Processing*, *Enviromental Information Processing*, *Cellular Processes*, *Organismal Systems*, *Human Diseases*, and *Drug Development*. This database includes general maps built using orthologs of genes as well as specific maps for each species, built jointly with a general map of metabolism. An example of a pathway map in KEGG

Category	Database	Entries
System Information	PATHWAY	467
	BRITE	158
	MODULE	632
Genomic information	ORTHOLOGY	18,307
	GENOME	3,510
	GENES	15,268,099
	SSDB	78,327,335,280
Chemical information	COMPOUND	17,339
	GLYCAN	10,987
	REACTION	9,776
	RPAIR	14,849
	RCLASS	2,945
	ENZYME	6,415
Health information	DRUG	10,118
	DISEASE	1,402
	ENVIRON	849

Table 1.5: **The Kyoto encyclopedia of Genes and Genomes (KEGG) database.** The databases categories and the number of KEGG entries in each of them are shown. (December,2014)

database is shown in Figure 1.6. Different nodes representing molecules and the different relationships between them are represented. Other databases which include larger amounts of curated data also exist, but KEGG is now the most used because it include a large number of processes and species (Table 1.5).

1.5 Pathway Analysis

Pathway analysis is now the first choice for the analysis of high-throughput data[29]. It allows insights into the underlying biology of differentially expressed genes to be gained, reduces complexity, and has an increased explanatory power.

Pathway analysis is understood as the set of approaches that identify pathways affected by a condition, correlating information about pathway knowledge

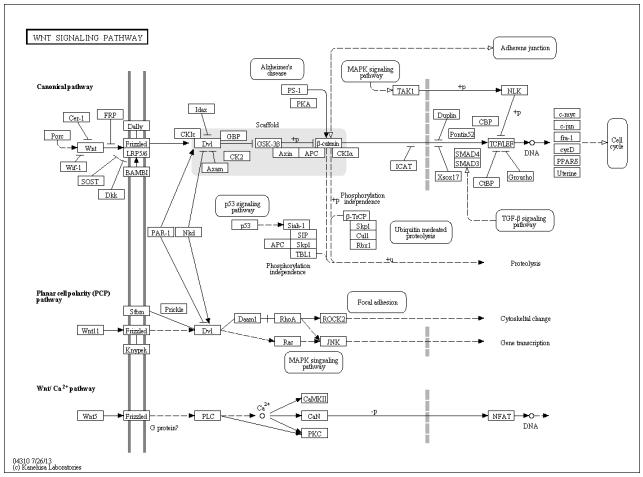


Figure 1.6: **KEGG maps example.** Reference pathway for WNT signaling in the KEGG database

with gene expression patterns. These approaches can be divided into three principal groups or generations according the type of analysis they perform[29].

First generation of methods are known as Over-Representation Analysis (ORA) approaches (competitive in other review papers[30]). Given a list of genes selected by a certain criteria, ORA approaches test if each selected pathway is enriched (over- or under-represented) by genes in the initial list. These methods are broadly used and several approaches following this philosophy exist. Nevertheless, this class of methods presents some limitations. Typically, the gene input list, which is necessary for these approaches, comes from a differential expression experiment where genes with p-values under a determined threshold are selected. Therefore, the results obtained are dependent of the chosen threshold which causes a loss of information because marginally less significant genes are missed. In addition, these approaches treat all genes in the list equally. Finally, these methods assume independence between pathways, even though they are interconnected because the limits of each pathway are determined only by the pathway database considered.

A new class of methods were developed to deal with most of ORA limitations. This second generation is referred to as Functional Class Scoring (FCS) approaches (self-contained methods in other review papers[30]). Given a list

of genes and their associated p-values (which come from a gene-level statistical test), FDR methods compute a running enrichment score for gene grouping. They are based on global and multivariate approaches that define a model on the whole gene set. This approaches are not dependent on any threshold and consider coordinate changes in gene expression, thus addressing the limitations of ORA methods. Nevertheless, these methods consider independence between pathways as ORA methods do. Also, they use changes in expression to rank genes according a given test, but they do not take into account the magnitude of the difference between genes belonging to the same class.

Finally, third generation include Pathway Topology (PT)-Based approaches. They are very similar to FCS methods, but take pathway topology (relationships between genes or proteins) into account to compute gene-level statistics. That is, they take advantage of the information available about relationships between genes that is included in pathway databases. These methods provide a impact factor (IF) measurement which considers the structure and dynamics of the pathway by incorporating; biological factors as differential expression measurements; and by considering the types of interactions; and the gene positions in the pathway network.

It has been argued that gene networks contain active subpathways (substructures in the pathway network) which are more relevant to gene expression, than all the pathway members taken together[31]. A new type of methods has recently emerged which is based on this idea. These methods use pathway topology in a similar way to PT-Based approaches, but the topology is not used to obtain a measurement of the behavior of the whole pathway, rather, it is used to dissect the pathway into subpathways. The final output these methods is an impact factor(IF) for each subpathway in a pathway; we will refer to these methods as Subpathway-based approaches (SP-based approaches). The first of these subpathway concept methods was `SubpathwayMiner`[32]. This method centers on the study of subpathways in the context of metabolic pathways. A subpathway is defined as a set of genes in which the distance between enzymes is less than a parameter called k (user-defined distance). To do this, each metabolic pathway is converted into an undirected graph with enzymes and nodes reducing the subpathway mining problem to a sub-graph mining problem. A p-value, based on hyper-geometric distribution, is then calculated to evaluate the significance of the enrichment for each structure (whole pathway or subpathway). This p-value takes the total number of genes on the genome, in the pathway, and the number of differentially expressed genes on the genome and in the pathway into account .

Other approaches based on subpathways have recently appeared. SP-Based

approaches agree than a subpathway is a subset of interconnected nodes in a pathway network, but two different subpathway concepts appear in the scientific literature. A subpathway can be considered as a subgraph or clique of the pathway graph, that is, a set of interconnected nodes, or it can be defined as an individual or linear path from a start-point (node without parent nodes) and an end-point (node without descendant nodes). Subpathway-based approaches can be divided in two groups based on this separation:

Clique based. There are two main approaches in this group: `DEGraph`[33] which is implemented as an R-package, and `clipper`[34] which is implemented as an R-package and as a web-tool called `GraphiteWeb`[35]. Both define subpathways as sets of interconnected genes. They also evaluate if differential expression pattern of genes in the subpathway agree with the different activations/inhibitions that are established between these genes. More specifically, `DEGraph` represents the pathway by a graph and a mean shift (vector of differences in the mean expression of all the genes in the pathway between two populations of interest); when the shift is coherent with the graph having low energy (regarding a particular energy function), it is evaluated. Finally, all the subpathways are tested one-by-one using a systematic approach to discover a non-homogeneous subgraph which results in a list of candidate subgraphs that represent the difference between the two conditions being studied. On the other hand, `clipper` converts the pathway network into an undirected cycle-free graph and applies two tests to this graph; one tests the behavior of the whole pathway using graphical Gaussian models for each condition, and the other identifies the relevant signal paths.

Linear path based. Most approaches based on this concept only describe the process without providing a tool [36, 37], however two approaches which provide a tool are highlighted in this section: the Topology Enrichment Analysis frameworK (TEAK)[38] and Differential Expression Analysis for Pathways (DEAP)[39]. `DEAP` is a python algorithm which examines any independent subpathway and calculates its differential expression by adding or subtracting all the downstream nodes with catalytic or inhibitory relationships. Then, `DEAP` calculates an absolute expression level value to determine which subpathway has the maximum value It also uses a self-contained approach which assesses the significance of each pathway. `TEAK` uses linear (from root to leaf) and nonlinear (represented by joining adjacent and overlapping feed-forward loop) subpathways. This ap-

proach calculates a score for each subpathway based on expression data and pathway structure, using a Bayesian Information Criterium (BIC) for context-specific data and Kullback-Leibler divergence for case/control data.

A summary of pathway analysis classification methods can be shown in Table 1.6 and Figure 1.7

name	availability	reference
ORA tools		
Onto-Express	Web	[40, 41]
GeneMapp	Standalone	[42]
GOminer	Standalone, Web	[43]
FatiGO	Web	[44]
GOstat	Web	[45]
FuncAssociate	Web	[46]
GOToolBox	Web	[47]
GeneMerge	Standalone, Web	[48]
GOEAST	Web	[49]
clueGO	Standalone	[50]
FuncSpec	Web	[51]
GARBAN	Web	[52]
GO:TermFinder	Standalone	[53]
WebGestalt	Web	[54]
WebGestalt2	Web	[55]
agriGO	Web	[56]
GOFFA	Standalone, Web	[57]
WEGO	Web	[58]
IPA	Web	–[5]
BayGO	R-package	[59]
ChipInfo	Web	[60]
PathMAPA	Standalone	[61]
ArrayXPath	Web	[62]
FCS tools		
GSEA	Standalone	[63]
GSEABase	R-package	[64]
sigPathway	R-package	[65]
Category	R-package	[66]
SAFE	R-package	[67]
GlobalTest	R-package	[68]
PCOT2	R-package	[69]
SAM-GS	Standalone	[70]
Catmap	Standalone	[71]
T-profiler	Web	[72]
FunCluster	Standalone	[73]

GeneTrail	Web	[74]
GAzer	Web	[75]
FatiSCAN	Web	[76]
PathwayMiner	Web	[77]
DAVID	Web	[78]
Pathway Processor	Standalone	[79]
PT-base tools		
ScorePAGE	No implementation available	[80]
Pathway-Express	Web	[81, 82]
SPIA	R-package	[83, 84]
NetGSA	No implementation available	[85]
BPAS	Standalone	[86]
Paradigm	Standalone	[87]
Subpathway-based tools		
SubpathwayMiner	R-package	[32]
DEgraph	R-package	[33]
Clipper	R-package	[34]
Graphite Web	Web	[35]
DEAP	Python algorithm	[39]
TEAK	Standalone	[38]
PATHiWAYS	Web	[88, 89]

Table 1.6: **Classification of pathway analysis methods**[29]. Several older methods are shown and a new group of subpathway methods (SP-based methods) are added. Pathway analysis methods are classified into four groups. Over-Representation Analysis (ORA) and Functional Class Scoring (FCS) are enrichment methods that do not take the structure of the network into account. Pathway topology (PT)-based methods do take the pathway structure, but use pathway information as a whole. Finally, SP-based methods take the internal substructures of the network that could lead to different functionalities into account.

[5] Ingenuity®Systems, www.ingenuity.com

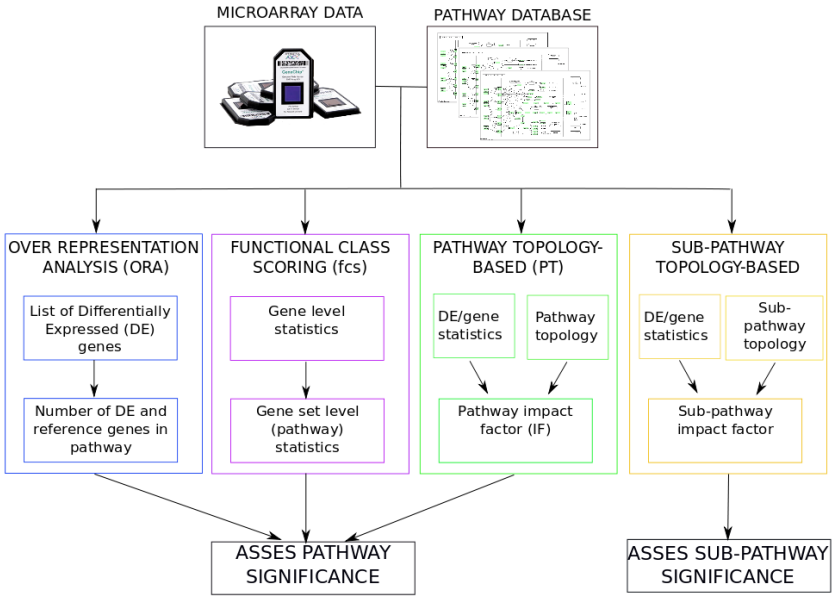


Figure 1.7: **Summary of pathway analysis methods classification**[29]. This methods are divided in four groups or generations. The first three groups ORA, FCS and PT-based (over-representation analysis, functional class scoring, and pathway topology respectively) are methods designed to asses the significance of the pathway, whereas the fourth one Subpathway-Based methods (SB-methods) appeared last year and aims to classify subpathways

Chapter

2 Motivation, objectives and contributions

"The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny...'"

— Isaac Asimov

2.1 Motivation

Cellular processes are often carried out via intricate systems of interacting genes and proteins, which are usually represented as biological pathways networks (Section 1.3). Biological pathways are annotated as the result of extensive effort by hundreds of researchers who manually contribute their experimental knowledge about a specific biological process by graphically representing it. Therefore, pathways are the best representation of biological experimentally validated knowledge for specific processes.

The impact of a specific condition on a pathway is broadly studied by several approaches which are divided into four different groups or generations (Section 1.5). Each generation was developed to overcome limitations of the methods in the previous generations, however most recent methods, SP-based approaches, still have some major limitations in terms of biological modeling and interpretation of subpathways.

The first group of SP-based methods, considers a pathway as a subgraph of the pathway graph. This definition implies that different stimulus (understood as the node that receives the external signal, i.e. start-point) and different effectors (understood as the node that provokes the cellular response, end-point) can be included in the same subpathway meaning that the functional interpretation of the results could be confusing. This limitation is overcome by the definition used for the second group of SB-methods, where a subpathway is defined as a linear path between a receptor (stimulus) and an effector. Therefore, different linear paths can be established between the same stimulus and effector because each of the linear paths represent a subpathway. Nevertheless, in reality, cellular functionality associated with the effector gene is a combination of the signals received through all of the linear paths between a stimulus and an effector. To overcome this limitation in the biological interpretation of subpathways, a new definition that combine both these ideas is proposed in this thesis. A subpathway is considered to be the set of linear paths going from a stimulus to an effector, where the combinatorial effects of all the linear paths between a stimulus and an effector are taken into account (Section 3.1.3).

Most of SP-based methods model information contained in pathway databases (specifically the KEGG database) without taking into account the biological meaning of the elements represented. Some of them consider an undirected graph, thus losing the directionality of the signal pass [32, 33]. Others fail to correctly modeling the information in the KEGG database [33, 35]. To overcome this limitation this thesis describes the process for modeling KEGG information to provide mathematical models that better reflect the biology

underlying the pathway models.

An important aim of Bioinformatics is to develop tools for analyzing of biological data which are accessible for non-expert users. Most of the tools provided by SP-based methods are implemented in the statistical programming language R, which drastically limits its use by inexperienced users, with the exception of `Graphite Web`[35] which implements `clipper` approach[34] and that provide a web-tool, and `TEAK`[38] that provides a standalone tool. In order to accomplish this aim, all the methods developed in this thesis were implemented in a web-tool called `PATHiWAYS`[89] (Section 4.7).

2.2 Objectives

The main objective of this thesis is to develop a methodology to analyze microarray expression data in a systems biology context. More specifically, we aimed to develop a method to dissect pathways in smaller substructures called subpathways, whilst retaining complete biological meaning: this tool was analyzed and experiment with two different conditions which allowed the functional differential behavior of subpathways to be assessed in a control-case experiment. This global objective can be broken down into smaller tasks:

- Create a database of models that describe the behavior of each probe set in different Affymetrix platforms. This database will be use as background data in the approach developed.

- Provide proper biologically oriented definitions of pathway objects and their relations to the KEGG database, thus obtaining models representing the underlying biology. Use these definitions to create a database of modeled pathways and subpathways that will be used as background structures for the analysis of expression data.

- Develop a statistical method for analyzing the differential activation behavior between two given samples in an expression dataset. This task includes the development of methodologies to estimate the probability of activation of each gene in the network and of signal transmission along the predefined subpathways, as well as the application of statistical tests to compare the differential behavior of each of them.

- Implement a tool which performs the methodology in order to bring it to the research community.

2.3 Contributions

2.3.1 Journal papers

1. **P. Sebastián-León**, J. Carbonell, F. Salavert, R. Sanchez, I. Medina, and J. Dopazo, "Inferring the functional effect of gene expression changes in signaling pathways", *Nucleic acids research*, p.gkt451, 2013

2. **P.Sebastian-Leon**, E. Vidal, P. Minguez, A. Conesa, S. Tarazona, A. Amadoz, C. Armero, F. Salavert, A. Vidal-Puig, D. Montaner, and J.Dopazo, "Understanding disease mechanisms with models of signaling pathway activities", *BMC systems biology*, vol.8, no.1, p.121, 2014

3. R.D. Hernansaiz-Ballesteros, F. Salavert, **P. Sebastián-León**, A. Aleman, I. Medina, and J. Dopazo, "Assessing the impact of mutations found in next generation sequencing data over human signalig pathways", *Nucleic acids research*, p.gk349, 2015

4. F. Eduati, LM. Mangravite, T. Wang, H. Tang, JC. Bare, R. Huang, T. Norman, M. Kellen, MP. Menden, J. Yang, X. Zhan, R. Zhong, G. Xiao, M. Xia, N. Abdo, O. Kosyk, S. Friend, G. Stolovitzky, A. Dearry, RR. Tice, A. Simeonov, I. Rusyn, FA. Wright, Y. Xie, S. Alaimo, A. Amadoz, M. Ammad-ud-din, CA. Azencott, J. Bacardit, P. Barron, E. Bernard, A. Beyer, S. Bin, A. van Bömmel, K. Borgwardt, AM. Brys, B. Caffrey, J. Chang, EG. Christodoulou, M. Clément-Ziza, T. Cohen, M. Cowherd, S. Demeyer, J. Dopazo, J. D Elhard, AO. Falcao, A. Ferro, DA. Friedenberg, R. Giugno, Y. Gong, JW. Gorospe, CA. Granville, D. Grimm, M. Heinig, RD. Hernansaiz, S. Hochreiter, LC. Huang, M. Huska, Y. Jiao, G. Klambauer, M. Kuhn, MB. Kursa, R. Kutum, N. Lazzarini, I. Lee, MKK. Leung, WK. Lim, C. Liu, FL. López, A. Mammana, A. Mayr, T. Michoel, M. Mongiovì, JD. Moore, R. Narasimhan, SO. Opiyo, G. Pandey, AL. Peabody, J. Perner, A. Pulvirenti, K. Rawlik, S. Reinhardt, CG. Riffle, D. Ruderfer, AJ. Sander, RS. Savage, E. Scornet, **P. Sebastian-Leon**, R. Sharan, CJ. Simon-Gabriel, V. Stoven, J. Sun, AL. Teixeira, A. Tenesa, JP. Vert, M. Vingron, T. Walter, S. Whalen, Z. Wiśniewska, Y. Wu, H. Xu, S. Zhang, J. Zhao, WJ. Zheng, D. Ziwei, and J. Saez-Rodriguez, "Prediction of human population responses to toxic compounds by a collaborative competition", *Nature biotechnology*, 2015.

5. A. Amadoz, **P. Sebastián-León**, E. Vidal, F. Salavert, and J. Dopazo, "Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity" *Scientific Reports [In Press]*, 2015

2.3.2 Conferences

July 2013 12th International Conference on Critical Assessment of Massive Data Analysis (CAMDA), Berlin (Germany), LONG PRESENTATION, Using probabilistic models of signaling pathways to predict in vivo drug activity.

November 2010 8th Meeting of the Valencian Network for Genomics and Proteomics, Valencia (Spain), POSTER, Dissecting signaling pathways to understand the consequences of gene expression changes.

October 2010 Xth Spanish Symposium on Bioinformatics (JBI2010), Málaga (Spain), POSTER, Dissecting signaling pathways to understand the consequences of gene expression changes.

2.3.3 Courses

March 2015 September 2014, March 2014 The Genomic of Gene Expression RNA-Seq Course, "Introduction to R and NGS Bioconductor packages", Teacher, 3.5h.

March 2015 September 2014, March 2014 The Genomic of Gene Expression RNA-Seq Course, "Introduction to Linux and command line", Teacher, 3.5h.

March 2011 VIII International Course of Massive Data Analysis, "Introduction to R and NGS Bioconductor packages", Teacher, 1.5h.

March 2011 VIII International Course of Massive Data Analysis, "Statistical Methods", Teacher, 3h.

June 2010 VII International Course of Massive Data Analysis, "Understanding genome-scale experiments at the light of protein-protein interactions", Assistant, 1h.

Chapter

3 Methods

"For every complex problem there's an answer that is clear, simple, and wrong"

— H.L. Mencken

In this chapter the methods developed and used in this thesis are explained. First, methods to generate background data are explained: pathway models (Section 3.1) and probe set expression models (Section 3.2). These steps are explained in the following sections and also appear in the method workflow (Figure 3.1). Starting from an expression dataset, the probability of activation of any probe set belonging to this dataset is calculated (Section 3.3) using the previously calculated mixtures of distribution(Section 3.2), summarized as the probability of activation of nodes or complexes (See section 3.4) which is propagated along the subpathways to estimate the probability of a subpathway being activated (Section 3.5). Once the probabilities of all the subpathways are compared, a statistical test is used to asses which ones are significantly different between both conditions (Section 3.6). Finally, significantly differentially activated subpathways are represented in a graph in order to put pathway map into a biological context (Section 3.7).

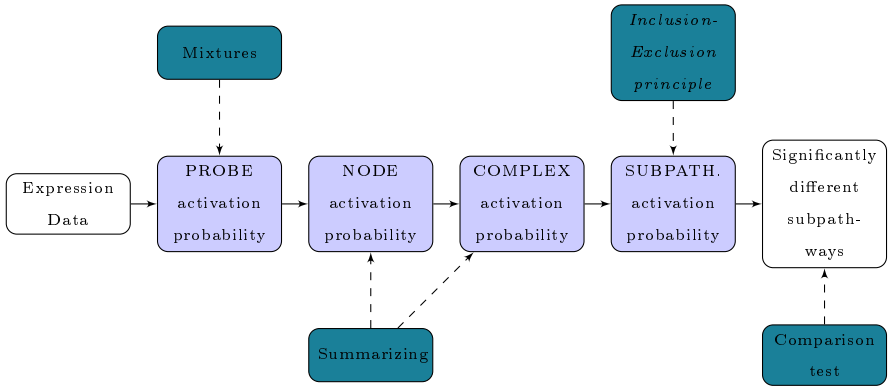


Figure 3.1: **Method's workflow** Expression data coming from a two-condition experiment is first converted into probe set activation probability using pre-calculated mixtures of probability distribution for each probe set on the platform. These probabilities are then summarized into node and complex probabilities using different rules, and this figures are subsequently transmitted along the subpathways using *Inclusion-Exclusion principle* to obtain the probability of activation for each subpathway in the pathway. Finally, a statistical test is applied in order to determined which subpathways changed to statistically significant degree between both conditions.

3.1 Pathway modeling

Information about the structure of the pathways was extracted from the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (See 1.4.2). This database is considered the foremost reference knowledge base for the integration and interpretation of large-scale molecular data sets, and is broadly used in pathways approaches. Information about nodes and their relationships is provided in KGML (KEGG Markup Language) format. Although this format is exclusive to KEGG, it is simple to parse because is very similar to the Extensible Markup Language (XML) format[1]

Nevertheless, this database does have some disadvantages. The information is computationally generated from manually defined information about it. This is hazy because the information about the figures provided by KEGG does not

[1] XML is a markup language that defines a set of rules for encoding documents. This format is both human and machine-readable and is broadly used in databases storage.

always match the information contained in the KGML file. Also, more than one gene is provided in a node without clarifying what the correct interpretation of this fact is (Section 3.4).

Metabolic and signaling pathways are different in nature and thus have to be modeled in a different way. Metabolic pathways produce metabolites via the action of enzymes (genes), while signaling pathways describe a signaling cascade from stimulus to receptors that performs a function in the cell. This thesis is focused on signaling pathways for two main reasons. First, signaling is closely related with cancer processes and, secondly, because mathematically modeling metabolic pathways, in the way proposed in this thesis[2], is not straightforward because many of reversible relationships exist that provoke cyclic behaviors. Also, the whole metabolic pathway usually is a cycle where stimulus and responses are not clearly defined.

3.1.1 KEGG Markup Language (KGML) structure

Information about nodes and relations in KEGG maps is provided in KGML format (Figure 3.2). KGML files start with a header indicating which version of KGML and KEGG the file belongs to. The first element in the file is called `pathway` and contains information about the pathway. Depending on this pathway we can find three elements: `entry` and, `relation` contain information about nodes and edges respectively, and `reaction` contains information about metabolic reactions and are out of the scope of this thesis.

KEGG `entry`

The `entry` element contains information about nodes in the pathway map (Table 3.1). The `name` slot indicates which genes are contained in the node and the `type` indicates which type the node is. Both are used to properly model the network (Section 3.1.2) and estimate the probability of the node being active (Section 3.4).

The `type` attribute has several values corresponding to different types of nodes (Table 3.2). Not all nodes defined in the KEGG database participate in signaling pathways. Specifically, only ortholog, gene, group, and compound nodes are considered in signaling pathways. Enzyme and reaction are nodes that act only in metabolic pathways, and map are nodes that represent other

[2]There exists several methods that model mathematically metabolic pathways, such Flux Balance Analysis (FBA)[90], but are mostly based in optimization problems

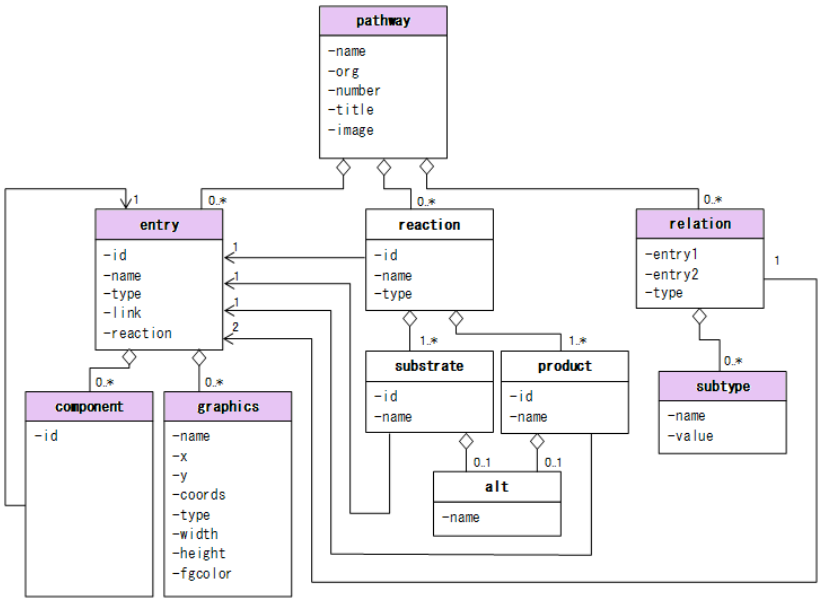


Figure 3.2: **KEGG Markup Language (KGML) elements**. Elements included in signaling pathways are highlighted in purple: `pathway`, `entry`, and `relation`. These elements contain general information about the pathway, nodes, and edges respectively.

attribute	explanation
id	the ID of this entry in the pathway map
name	the KEGGID of this entry
type	the type of this entry
link	the resource location of the information about this entry
reaction	the KEGGID of corresponding reaction
graphics	graphical information

Table 3.1: **KEGG `entry` information**. The information used in this thesis is highlighted in bold.

pathways. These nodes are not considered because they lack of relations associated with map nodes in KEGG database.

type	explanation
ortholog	the node is a KO (ortholog group)
enzyme	the node is an enzyme
reaction	the node is a reaction
gene	the node is a gene product (usually a protein)
group	the node is a complex of gene products (usually a protein complex)
compound	the node is a chemical compound (including a glycan)
map	the node is a linked pathway map

Table 3.2: **Node types in the KEGG database**. Node types included in signaling pathways are in bold

It is noteworthy that both the genes and ortholog nodes, can contain more than one gene/ortholog inside. That is, a node does not always represent a gene, but it can represent a set of genes. Interestingly, there are some analysis approaches that model this situation as a set of genes that works independently between them[33, 34]. However, this is not the most realistic assumption. KEGG maps are static and their components are shared for different tissues and developmental stages. Most of the genes contained in a node belong to the same family, and share their functionality, but they are expressed in different tissues or in at different developmental stages. For example, in the *VEFG signaling pathway* (Figure 3.3a) we can find a node, labeled as SPK that includes two genes: sphingosine kinase 1 and 2 (SPHK1 and SPHK2). According to UniProtKB/Swiss-Prot database[91] the function of both is to catalyze the phosphorylation of sphingosine to form sphingosine 1-phosphate (SPP)[92, 93], but they have a different subcellular location. SPHK1 is located in the cytoplasm[94] and nucleus, and SPHK2 has two isoforms, the first located in the cytoplasm and membrane (isoform 1) and the second in lysosome membranes (isoform 2)[95, 93]. Therefore, we consider them as a simple node, where the genes contained within it act alternatively to pass the signal along the path.

Group nodes represent a different situation. This label indicates that more than one node acts cooperatively. That is, all nodes must be activated to mantain its integrity. These nodes are considered to be complex nodes and their behavior is modeled in a different way than simple nodes. Other Subpathway-based methods consider them as separate nodes, but this is not ideal because nodes must act together to pass on signals, as is shown in Figure 3.3b. In the *PPAR signaling pathway*, surrounding nodes are indicated as groups in the

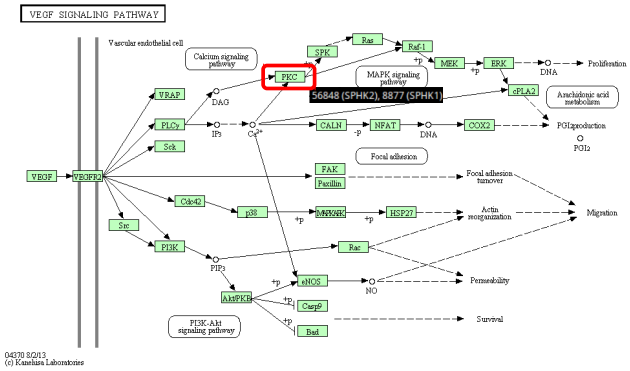
KGML; these groups are described as complexes in GeneCards[96] (one of the most curated databases for gene information) and indicates that PPARs form heterodimers with retinoid X receptors (RXRs) and that these heterodimers regulate transcription of various genes[3].

In KEGG maps, rectangles represent gene products, which are usually proteins, but can also be RNA or other complexes; circles are other types of molecules, usually chemical compounds, while round rectangles represent links to other pathways. This information is contained in the `graphics` element (Table 3.1) as well as in the position of each node and is used to define the layout of the graphical output of the pathway (Section 3.7).

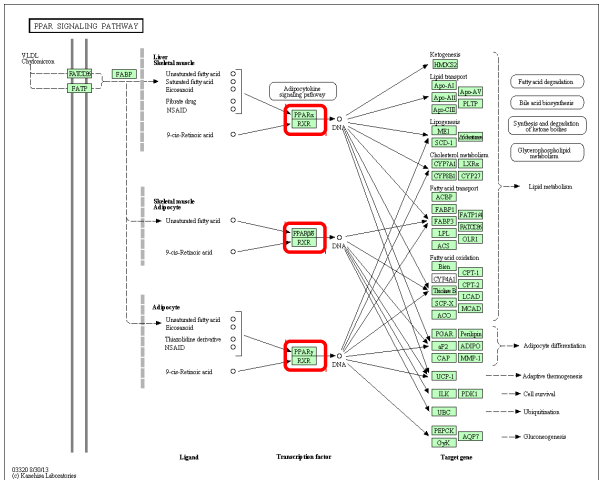
attribute	explanation
name	the label of this graphics object
x	the X axis position of this graphics object
y	the Y axis position of this graphics object
coords	the polyline coordinates
type	the shape of this graphics object
width	the width of this graphics object
height	the height of this graphics object
fgcolor	the foreground color used by this graphics object
bgcolor	the background color used by this graphics object

Table 3.3: **KEGG `graphics` information**. Information used for the development of this thesis is highlighted in bold

[3] `http://www.ncbi.nlm.nih.gov/gene?db=gene&cmd=Retrieve&dopt=full_report&list_uids=5468`



(a) *VEFG signaling pathway*



(b) *PPAR signaling pathway*

Figure 3.3: **Examples of simple and complex nodes**. **(a)** A node labeled as SPK, in *VEFG signaling pathway*, includes information about two genes.**(b)** Surrounded nodes in *PPAR signaling pathway* nodes are indicated as groups in the KGML file

KEGG `relation`

The `relation` elements contain information about the relationships that are established for the node in the pathway map (Table 3.5).

attribute	explanation
entry1	the first (from) entry that defines the relation
entry2	the second (to) entry that defines the relation
type	the type of the relation
subtype	the class of the relation

Table 3.4: **Relation information in KEGG database.**

The `type` attribute indicate the class of the elements that take part in the relation (Table 3.5) and the `name` attribute in the `subtype` slot indicates the class of relation which has been established between the nodes (Table 3.6). There is not an single `subtype` for each relation. For example a relation can be a phosphorylation event that provokes inhibition of the upstream protein, thus the `relation` element has two subtypes: phosphorylation and inhibition.

value	explanation
ECrel	enzyme-enzyme relation, indicating two enzymes catalyzing successive reaction steps
PPrel	protein-protein interaction, such as binding and modification
GErel	gene expression interaction, indicating a relation between a transcription factor and target gene product
PCrel	protein-compound interaction
maplink	link to another map

Table 3.5: **Relation `types` in KEGG database.** Relation types included in signaling pathways are highlighted in bold

Not all relationships defined by KEGG are included in signaling pathways. Relationships between nodes can be defined in terms of the activation/deactivation effect on the upstream protein, i.e. its capacity to transmit the signal along the path:

compound A compound binds another element to activate the upstream protein.

value	explanation
compound	shared between two successive reactions (ECrel) or intermediate of two interacting proteins (PPrel)
hidden compound	shared between two successive reactions but not displayed in the pathway map
activation	positive effects which may be associated with molecular information
inhibition	negative effects which may be associated with molecular information
expression	interaction via DNA binding
binding/association	proteins association
dissociation	proteins disassociation
phosphorylation	
dephosphorylation	
glycosylation	molecular events
ubiquitination	
methylation	

Table 3.6: **Relation names in the KEGG database.** Relation names included in signaling pathways are highlighted in bold

activation A downstream protein activates upstream protein.

inhibition Activation of the upstream protein is blocked.

expression Expression of the upstream protein is activated.

binding/association Both proteins bind to each other to regulate another protein.

dissociation The separation of two bound proteins resulting of the activation of the upstream protein.

phosphorylation The upstream protein is activated/inhibited by the addition of a phosphate group.

dephosphorylation The upstream protein is activated/inhibited by the loss of a phosphate group.

ubiquitination The upstream protein is degraded by the action of a ubiquitin protein.

In KEGG pathway maps, the edges between rectangles represent functional interactions and they can be directed, directed with a $+p$ (phosphorylation), directed with a $-p$ (dephosphorylation, or directed with a bar at the end (inhibition).

3.1.2 KEGG signaling pathways modeling

First, KGML files were downloaded from the KEGG database (Release 66.0, April 1, 2013). Information about the components of the network and their relationships were extracted from KGML files using a PERL script.

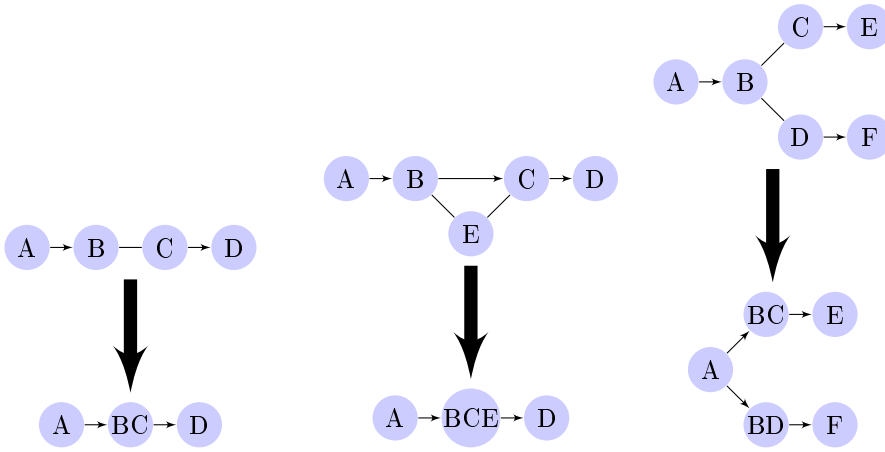
The relations, except binding/association relationships, were simplified into two possible events: activation and inhibition. Relations that provoke the activation of the upstream protein, allowing the signal to pass through the subpathway are considered to be activations. KEGG allows more than one relation between nodes (Section 3.1.1), that is, a relation between two nodes can be indicated as phosphorylation and inhibition together, and in this case is a phosphorylation that produces an inhibition of the upstream protein. In general, all relations are considered to be activations unless it is indicated as an inhibition or ubiquitination in the KGML file. Ubiquitination is a relationship where a ubiquitin protein is attached to another protein to provoke its degradation which consequently causes an interruption of the signal passing through the subpathway. Therefore, inhibition and ubiquitination are both considered to be inhibitions, and other types or relations are considered to be activations.

The binding/association relation is different to the others because it does not indicate any modification in the activation of adjacent proteins, but it does indicate that two proteins must bind together in order for the signal to pass along the path. This situation is similar to group nodes (Section 3.1.1), and so it is modeled in the same way. Nevertheless, binding/association modeling is not as evident as it is for group nodes, and consequently several proteins which are related by binding/association are joined together with other nodes in the network for other relations. Therefore, we had to establish some rules to specify how these complex nodes bind (Figure 3.4):

- Nodes connected only by a binding/association relation are considered to be complex nodes (Figure 3.4a).

- Nodes connected by both, binding/association and another relation (activation or inhibition), are joined into a complex node (Figure 3.4b).

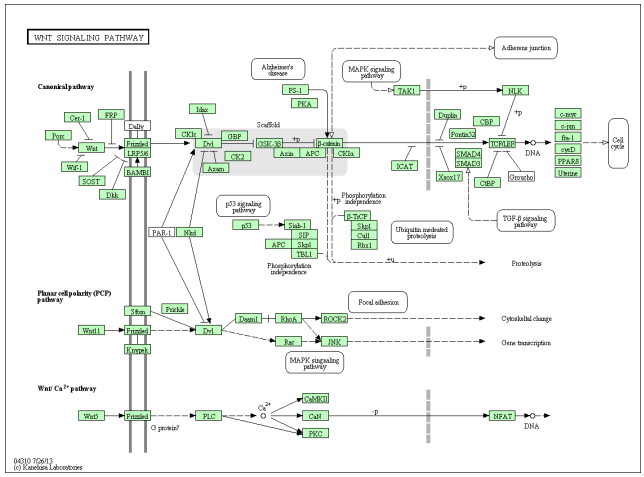
- Nodes connected by a binding/association with nodes that provoke a bifurcation are duplicated and are considered once in each bifurcation subpathway (Figure 3.4c).



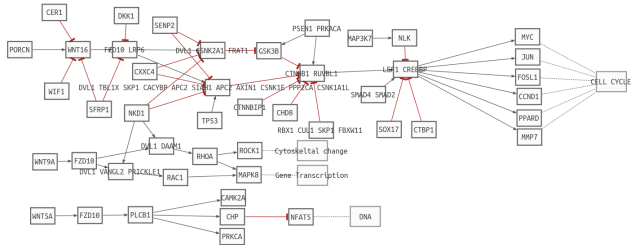
(a) Rule 1 (b) Rule 2 (c) Rule 3

Figure 3.4: **Rules for complex nodes set.** Two types of relationships are considered in this figure: Activation/Inhibition (→), and binding/association (—)

Modeling a binding/association relation in this way is one of the strengths of this approach. Most SP-based methods consider this relationship as a non-directed edge and consider the nodes that are related by it as independent, but this modeling does not reflect the biology underlying these relationships. In Figure (3.5) the *WNT signaling pathway* is modeled using the rules explained in this section. Two types of nodes are considered: simple nodes (white) and complex nodes (gray). The, relations are simplified into two types: activation (black) and inhibition (red). Nodes in the shaded rectangle are joined by binding/association relationships according to the KGML file and are indicated as scaffold proteins in the KEGG map (Figure 3.5a). Once the modeling rules are applied, some complex nodes are created (Figure 3.5b). It is of note that not all the shaded nodes are joined together, in fact DVL1 node is duplicated allowing the signal to pass through two subpathways.



(a) KEGG map



(b) Modeled network

Figure 3.5: **An example of *WNT signaling pathway* modeling**. **(a)** The KEGG map for the *WNT signaling pathway* **(b)** The modeled network for the *WNT signaling pathway*. Only two types of nodes (simple/complex) and relations (activation/inhibition) are considered.

3.1.3 Subpathway extraction

Once a pathway is modeled as indicated in section 3.1.2, it can be represented by a graph where genes are nodes and their interactions are edges, although to define a subpathway, some concepts about graph theory must be defined[97]:

Graph A graph $G = (V, E)$ in two sets $V = \{v_1, ..., v_N\}$ and $E = \{e_1, ..., e_M\} =$

$\{(v_i, v_j/(v_i, v_j) \in VxV)\}$ where:

- The elements of V are called vertices (or nodes).
- The elements of E are called edges.
- Each edge has a set of one or two vertices associated with it, which are called its endpoints. An edge is said to join its endpoints.

Adjacency A vertex v_i is adjacent to vertex v_j if they are joined by an edge.

Directionality A directed edge (or arc) is an edge $e_k = (v_i, v_j)$, in which one of its endpoints is designated as the tail, and the other as the head (a directed edge is directed from its tail to its head). A digraph (or directed graph) is a graph where each of the edges is directed.

Degree The degree (or valence) of a vertex v in a graph G is the number of proper edges incident to v.The in-degree of a vertex v in a digraph is the number of arcs directed to v; and the out-degree of a vertex v is the number of arcs directed from v.

Based on these concepts, we define an **entry node** (stimulus) as a node showing zero in-degree, except if it acts as an inhibitor. Similarly, an **exit node** (response) is a zero out-degree node. A set of adjacent nodes between two nodes is considered to be a **path**. Finally, a stimulus-response **subpathway** is defined as the set of paths between an entry and an exit node. These stimulus-response subpathways represent the way in which the signal is transmitted along the pathway network. This transmission may occur along a single path or via multiple paths between entry and response nodes (Figure 3.6). Each subpathway, as defined above, is an "atomic" functionality within the pathway.

To extract all the subpathways from a pathway, we developed an algorithm in R that calculates all possible paths between an entry and an exit node. This algorithm automatically discards looped paths and groups all the paths which share the same entry and exit node in a subpathway.

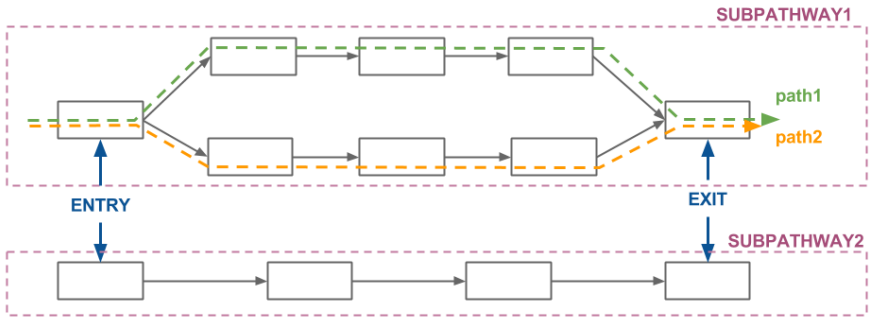


Figure 3.6: **Subpathway elements** Two different entries and two different exits can be observed on this sample network, as well as two subpathways defined between them. The first subpathway comprises two paths (two possible routes that the signal can take from entry to exit).

3.2 Probeset expression distribution modeling

A comparable measurement criteria is needed to assess the capacity of each node to transmit a signal through a subpathway, however, the level of expression is not comparable itself (Section 1.2.2). In 2007, Efroni[7] proposed a method to estimate the probability of a node's activity by modeling the expression level of a gene as a mixture of two distributions: one associated with the active state of the gene and the other its inactive state. These mixtures will serve as a template of the probability of activation of each probe set given its expression value.

Consequently, we decided to use the activation probability for each probe set as our measurement of the probe set's activity. Conceptually it is understood that the classification of genes into alternative states is a simplification of much more complex complexity patterns of gene behavior and action[18]. Therefore, the expression of each probe set was modeled as follows:

$$p(x) = \pi_1 p_1 + \pi_2 p_2$$

where p_i are the distributions of the active and inactive states and π_i are the estimated mixture coefficients.

3.2.1 Data preprocessing

Each Affymetrix DNA microarray measures a different set of probe sets, and the range of expression common to each probe set is also very different. Therefore the model of each probe set is dependent on the platform used to measure it. Six expression microarray platforms were considered in this thesis, three human (*Homo Sapiens*), two mouse (*Mus Musculus*) and one rat (*Rattus norvegicus*). The modeled platform and its abbreviations are shown in Table 3.7.

In order to estimate the distribution of each probe set in each platform, data from the Gene Expression Omnibus (GEO)[25] database was downloaded. The *GEOmetadb* R-package from Bioconductor[98] was used to render a total of 43,913 arrays for the selected platforms. These arrays were hybridized for every type of human, mouse, and rat samples. Data collection covered an ample spectrum of biological conditions including different tissues, diseases, male and female individuals, as well as different cell lines.

For each set of arrays (one for each platform), raw data (.CEL files) were downloaded, normalized in batches of 100 (because of memory size limitations) using the RMA function from *affy*[99] Bioconductor R-package and finally all

Platform	Abbreviation
Affymetrix Human Genome U133 Plus 2.0 Array	HGU133Plus2
Affymetrix Human Genome U133A Array	HGU133A
Affymetrix Mouse Gene 1.0 ST Array	Mouse1st
Affymetrix Mouse Genome 430 2.0 Array	MG430_2
Affymetrix Human Genome U133A 2.0 Array	HGU133A_2
Affymetrix Rat Genome 230 2.0 Array	RG230_2

Table 3.7: **Selected platforms and their abbreviations**. Three platforms for *Homo Sapiens* (HGU133Plus2, HGU133A and HGU133A_2), two for *Mus Musculus* (Mouse1st and MG430_2) and one for *Rattus norvegicus* (RG230_2) are used

the batches were rescaled together using the *quantile* method in the *limma*[100] Bioconductor R-package.

3.2.2 Estimation of mixture parameters

Mixtures modeling was performed using *mixdist*[101] CRAN (Comprehensive R Archive Network) R-package, more specifically, the *MIX* function developed by MacDonald in 1988[102]which analyzes histograms as mixtures of statistical distributions. This function is used to find a set of overlapping component distributions that gives the best fit to the histogram using the maximum likelihood. *MIX* assumes that:

- Components can be described by either normal, log-normal, or gamma probability distributions
- Data are grouped, in the form of numbers of observations over successive intervals

The *MIX* method also performs a goodness-of-fit test depending on the chi-square approximation to the likelihood ratio statistic[103]. The degrees of freedom are computed as the number of grouping intervals minus 1 minus the number of parameters estimated.

The *MIX* function returns an estimated value for each of the parameters: the mean and the variance of each distribution, mixture coefficients and the goodness-of-fit p-value of the mixture. Mean and variance are not the exact

the parameters, but are a linear combination of them. To estimate both the parameters of the distributions and the mixture coefficients, the following steps were performed for each probe set:

1. Calculate the two highest picks in the expression distribution of the probe set., which are used as the initial point in the calculation of the mixture.

2. Calculate the mixture that best fits the distribution of the expression. This step is done for gamma, normal and log-normal distributions.

3. Use the goodness-of-fit p-value to choose the distribution that best fits the distribution between gamma, normal and log-normalization. The distribution with the lowesst p-value is selected. This p-value can be not significant in both cases, because it is only used to compare the mixture model that best fits the expression distribution, although this distribution does not fit significantly.

4. Estimate the parameters of the optimal distribution from the estimated mean ($\hat{\mu}$) and variance ($\hat{\sigma}^2$). If the optimal distribution is normal, the parameters match, but if it is gamma or log-normal, the following parametrization was used:

$$\mathrm{Ga}(k_{GA} = \frac{\hat{\mu}}{\theta_{GA}}, \theta_{GA} = \frac{\hat{\sigma}^2}{\hat{\mu}})$$

$$\mathrm{LN}(\mu_{LN} = 2\log(\hat{\mu}) - \hat{\sigma}_{LN}, \hat{\sigma}_{LN} = \sqrt{\log(\hat{\sigma}^2 + 1)})$$

5. An .RData object with the following information was used: the selected distribution (best fit) of the components of the mixture (p_1 and p_2), the estimated mixture coefficients for each distribution (π_1 and π_2) and their estimated scale and shape parameters.

3.3 Probe set activation probability

Once a mixture distribution is estimated, the probability of a probe set being in the active distribution can be estimated based on its expression value[7]. According to *Bayes' Theorem*, the activation probability of a probe set given a specific expression value (x) is:

$$P(Active|x) = \frac{p(x|Active)P(Active)}{p(x)}$$

where, according to *Law of total probability*,

$$p(x) = p(x|Inactive)P(Inactive) + p(x|Active)P(Active)$$

In terms of the mixture distribution, it can be written as:

$$P(Activo|x) = \frac{\pi_1 p_1(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)}$$

where,

- $\pi_0 = P(Inactive)$ y $\pi_1 = P(Active)$ are the mixture coefficients estimated in the previous section.
- $p_0(x) = p(x|Inactive)$ y $p_1(x) = p(x|Active)$ are the mixture components estimated in the previous section.

3.4 Node activation probability

There are two types of nodes in the modeled network: simple and complex nodes (Section 3.1.1) that model different biological situations. Simple nodes can comprise more than one gene and indicate an OR relation between their components, while complex nodes are composed of simple nodes joined by an AND relation.

There are three types of simple nodes in KEGG signaling pathways: compound, ortholog, and gene. Orthologs and compounds are not measured by expression microarrays, therefore th probability of activation of both of them is defined in one single term[4]. Gene nodes can contain more than one gene (Section 3.1.1) and any gene can be mapped by more than one probe set (Section 1.2.1). Expression level is measured at the probe set level, so the set of genes contained in a node is converted into a set of probe sets taking into account that a probe set mapping more than one gene is removed from the analysis where confused measures could be included, but is kept where all the genes mapped by it are included in the same gene. The activation probability of probe sets mapping to a node can be summarized using different statistics: the mean, median, maximum, minimum, and 90th, 95th and 99th percentiles. All of these statistics should be valid, but according to the definition of simple nodes it is reasonable to use the most activated gene as a proxy for node

[4]These nodes are maintained in the KEGG structure for future implementations which incorporate their measurements

activation, because genes included in this type of node are signal transmission alternatives. Therefore, in this thesis, the activation of a simple node is summarized using the 90th percentile of the probe sets mapping to it; the maximum was discarded in order to avoid extreme values.

On the other hand, complex nodes represent a set of nodes acting cooperatively (e.g. a scaffold protein), therefore, nodes comprising it are essential for its integrity (signal transmission). Consequently, the activation probability of a complex node is summarized by the minimum activation probability of the simple nodes it includes.

3.5 Propagation of probabilities in the network

Once the activation probability of each node is estimated, must be propagated along the subpathway in order to estimate its activation probability (Figure 3.7). This is done taking into account that:

- Two nodes joined by an activation ($A \rightarrow B$) will transmit the signal if both are active, so the probability of signal transmission is: $P(A\ active) * P(B\ active)$
- Two nodes joined by an inhibition ($A \dashv B$) will transmit the signal only if the inhibitor is not active, so signal transmission probability is: $P(A\ inactive) * P(B\ active) = [1 - P(A\ active)] * P(B\ active)$
- In subpathways formed by more than one signaling path, the signal can be transmitted by each of them, so the *inclusion-exclusion principle*[104] is used to estimate the signal transmission probability.

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \bigcap A_j) + \\ = \sum_{i<j<k} P(A_i \bigcap A_j \bigcap A_k) + \cdots + (-1)^{n-1} P(\bigcap_{i=1}^{n} A_i)$$

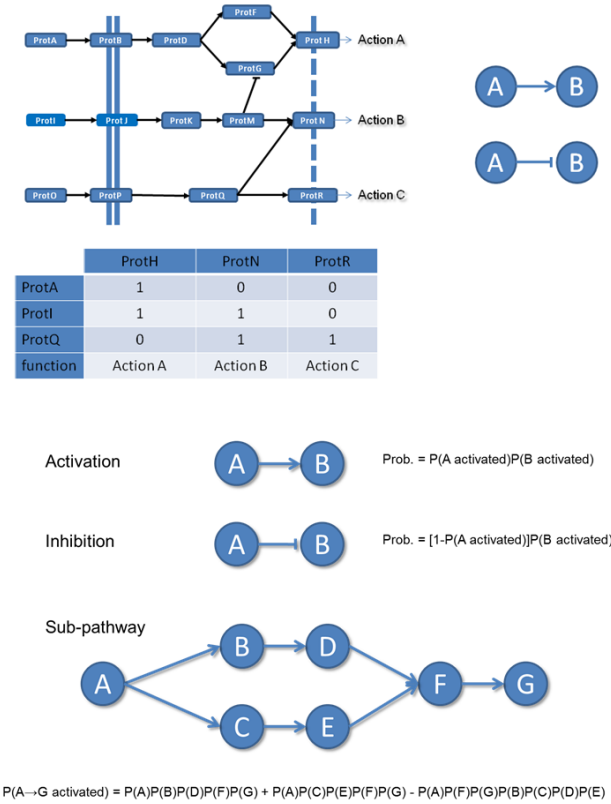


Figure 3.7: **An example pathway with three possible inputs and three possible outputs**. Any of the output proteins triggers a different pathway functionality. The connection matrix shows how the input and output can be connected through five different subpathways. Any subpathway can be traversed by different paths. For example, the subpathway connecting protein A (ProtA) to protein H (ProtH) can be traversed by two different paths, represented by two node sequences (that, for the sake of the simplicity, are equivalent to proteins in this sample): ProtA, ProtB, ProtD, ProtF, ProtH or ProtA, ProtB, ProtD, ProtG, ProtH. On the right, the two interactions between proteins are shown: the top right represents activation, with an arrowhead line, and bottom right represents repression with a line with no arrow. In the lower part there is an example to illustrate the way in which the probability of subpathway activation can be calculated from the combined activation of the corresponding nodes. From top to bottom: i) probability of the signal transmission through an activation action, ii) probability of the signal transmission through a repression action, iii) an example of a simple bifurcating sub-pathway, and iv) the probability of signal transmission along this pathway.

3.6 Comparison in a control-case experiment

The probability of activation is not informative by itself, because biological information is not easily extracted from a subpathway activation probability. Subpathway activation probabilities are, by definition, small, considering that they are the result of multiply several probabilities belonging to the [0,1] interval. The larger the number of nodes belonging to a subpathway, the smaller the activation probability. Therefore, these subpathway activation probabilities are typically used for making comparisons. The biological meaning of having a significantly lower or higher activation probability value in a condition with respect to another is clear because significant changes in the signal transmission probabilities in a functional subpathway are a cellular response to a change in conditions.

Our main concern using this approach is to assess the differential activation probability of every subpathway in each pathway, given the sample status (disease or control). This data for a given experiment is a matrix which columns indicate samples for both groups to compare and rows indicate all the possible subpathways in the selected pathways. Matrix values represent the activation probability for each subpathway on each sample.

Two possible strategies can be applied to solve this problem: parametric and non-parametric statistics. Parametric statistics assumes that data comes from a known distribution, and inferences are made about the parameters of the underlying distribution. On the other hand, non-parametric statistics do not rely on data belonging to a particular distribution. The Student t-test and multilevel modeling are used in parametric statistics, and Wilcoxon test is used for non-parametric statistics.

3.6.1 Parametric testing

Two different options were considered for parametric testing: the student t-test and multilevel modeling.

The most well-known parametric test is Student t-test[105], and so it was tested first, although the subpathway probabilities are clearly not normally distributed. This group of tests refers to any statistical hypothesis in which the test statistic follows a Student's t distribution if the null hypothesis is supported, although it is most commonly applied when the test statistic follows a normal distribution, given a known value of scaling term. A False Discovery Rate (FDR) control[106] is then applied to correct the p-values obtained for each subpathway in the pathway, which can be used to assess which sub-

pathways show a significant difference in activation between conditions. The t-statistic estimation is used to assess which condition is more activated in each significant subpathway.

For multilevel modeling, an average activation probability for each pathway was supposed, taking into account possible differences because of the status and allowing inter-subpathway (intra-pathway) variation via a random subpathway random effect. This model uses a multilevel varying-intercept, varying-slope regression with activation probability (p_i) as an outcome, sample status (s_i) as a predictor, and subpathway (j) as a group factor. The parameters associated with the status indicator variable(μ_β for the whole pathway, β_j for each subpathay) have a simple and useful interpretation: they stand for the average difference in activation probability between the disease and control samples. This model can be viewed as an extension to the classical comparison of proportions. In the standard approach, activation of each subpathway would be compared independently of the rest of the sample and the variability corrected for multiple comparisons. In the proposed model, subpathway and pathway information are used together to estimate both the location and variability of the parameters. Information is shared between all subpathways that make up a pathway by shrinking each path estimator from the standard analysis value (unpooled, considering the subpathways as independent) to the pathway overall average value (pooled, considering all paths together). The extent of the shrinkage depends on the variability of each path: the more variable (less consistent, less reliably informative) a path is, the more its estimation is moved towards the overall average.

This allows multiple comparison problems to be avoided[107]. In this case, we did not use a logit transformation on the activation probabilities, mainly for the two following reasons:

1. The parameter interpretation is less intuitive (odds ratio vs. absolute probability differences)

2. The instability of estimations in the presence of 'outliers' (extreme probabilities) are a current hurdle in the development of more sophisticated multilevel modeling techniques.

3.6.2 Non-parametric testing

Non-parametric testing was performed using the Wilcoxon test[108] implemented in R (*wilcox.test* function) to compare the difference in the activation probability of each individual subpathway in the pathway. A False Discovery

Rate (FDR) control[106] was then applied in the same way as in the t-test case. The p-value obtained was used to asses any subpathways with a significant difference in activation between conditions. Estimation of the location parameter is used to asses which condition is more activated in each significant sub-pathway.

3.7 Graphical interpretation

Results can be shown in a table indicating each sub-pathway and their corresponding p-values, FDR corrected p-value and a label 'UP/DOWN' indicating if the activation probability is higher in the first or in the second group respectively. For a better understanding of these results in a pathway context, significant results are represented in a KEGG-like map. This representation is obtained using the R-package igraph[109] in a first approach[88, 89] and is followed up by using CellMaps (`http://cellmaps.babelomics.org/`). Nodes belonging to subpathways which are significantly more activated in the first condition appear in blue, while those which were significantly more activated in the second condition appear in red. Nodes belonging to several subpathways with different behavior appear in purple (Figure 3.8).

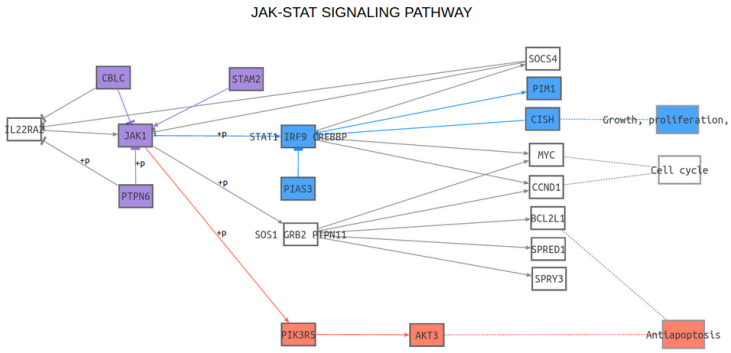


Figure 3.8: **Example of a graphical representation of the results[5] in the *JAK-STAT signaling pathway***. Nodes belonging to significantly more activated subpathways in the first condition appear in blue, and ones which weresignificantly more activated in the second condition appear in red. Nodes belonging to several sub-pathways with different behaviors appear in purple.

Chapter

4 Data and results

"Look deep into nature, and then you will understand everything better."

— Albert Einstein

This chapter starts with a summary of the data used to generate the background data: modeled pathways (Section 4.1) and probe sets' mixtures (Section 4.2). This is followed by an analysis to validate the methodology (Section 4.3). I go on to describe some results obtaining by applying the developed approach to real data (Section 4.4) in order to probe the utility of the methodology. The method is them compared with other existing methods using a colorectal cancer dataset (Section 4.5). PATHiWAYS approach was also used to take part in the CAMDA'2013 challenge and the results are included here (Section 4.6). Finally, the repackaging of the approach into an user-friendly web-tool is explained in section 4.7

4.1 Selected pathways

Not all signaling pathways in the KEGG database could be modeled for three main reasons:

- There were pathways which nearly formed a complete loop. As indicated in section 3.1.3, looped paths are discarted, because these pathways cannot be modeled.
- Pathways containing subpathways with a huge number of paths. In this case, it is not possible to compute the Inclusion-Exclusion principle formula within a reasonable period of time, and so these pathways are not modeled.
- Nodes where the information available for different species is inconsistent or not suficient to extract comprehensive subpathways.

According to these criteria, 27 human (*Homo Sapiens*) KEGG signaling pathways were selected and modeled as previously described (section 3.1.3). 18 signaling pathways for mouse (*Mus musculus*) and 22 for rat (*Rattus Novergicus*) were subsequenly included as they were undertaken for different studies. The selected pathways for each species are described in Table 4.1.

The number of subpathways in each pathway varied a lot (Figure 4.1), from four subpathways in the *Hedgehog signaling pathway* to 179 subpathways in the *Cytokine-Cytokine receptor interaction* pathway.

The distribution of the subpathway elements also differeed depending on the pathway (Figure 4.2). The number of nodes (a) can be very variable, as is the case for the *Apoptosis* or *Adipocytokine signaling pathway*; or it can be constant as for example in the *PPAR signaling pathway*, where the subpathways are composed of three nodes or *Neuroactive ligand-receptor interaction*, *Antigen processing and presentation*, and *Cytokine-Cytokine receptor interaction* pathways, whose subpathways comprise two nodes. Nodes can act as activators or inhibitors: there are some pathways without inhibitors (e.g the *PPAR signaling pathway*) and some with more inhibitors than activators per subpathway (e.g. the *WNT signaling pathway*). Finally, there are pathways composed of linear subpathways (e.g. the *Tight junction*) and others with a large number of paths per pathway (e.g. the *ERBB signaling pathway*).

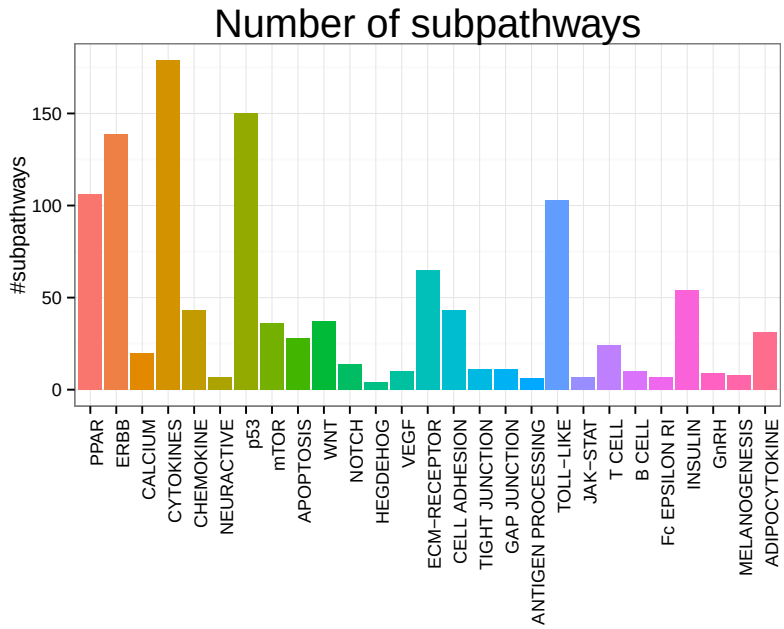


Figure 4.1: **Number of subpathways per pathway**. There are pathways with a large number of subpathways (e.g. the *Cytokine-Cytokine receptor interaction*) as well as pathways composed by very few subpathways (e.g. the *Hedgehog signaling pathway*)

Environmental Information Processing		
KEGG subgroup	Pathway name	Species
Signal Transduction	ErbB signaling pathway	h,m,r
	Wnt signaling pathway	h,m,r
	Notch Signaling pathway	h,m,r
	Jak-STAT signaling pathway	h,m,r
	Calcium Signaling pathway	h,m,r
	VEGF signaling pathway	h,m,r
	Hedgehog signaling pathway	h,r
	mTOR signaling pathway	h
Signaling molecules and interaction	Neuroactive ligand-receptor interaction	h,m
	Cell adhesion molecules (CAMs)	h,m,r
	Cytokine-cytokine receptor interaction	h,r
	ECM-receptor interaction	h,r
Cellular process		
KEGG subgroup	Pathway name	Species
Cell growth and death	Apoptosis	h,m,r
	p53 signaling pathway	h,r
Cell communication	Gap junction	h,m,r
	Tight junction	h,r
Endocrine system	Insulin signaling pathway	h,m
	Adipocytokine signaling pathway	h,m,r
	PPAR signaling pathway	h,m,r
	GnRH signaling pathway	h,m,r
	Melanogenesis	h,m,r
Immune system	Toll-like receptor signaling pathway	h
	B cell receptor signaling pathway	h,m,r
	T cell receptor signaling pathway	h,r
	Fc epsilon RI signaling pathway	h,m,r
	Antigen processing and presentation	h,m,r
	Chemokine signaling pathway	h

Table 4.1: **Selected pathways.** Classification of the selected pathways into KEGG categories and according tothe species in which they are modeled. h=human, m=mouse, r=rat

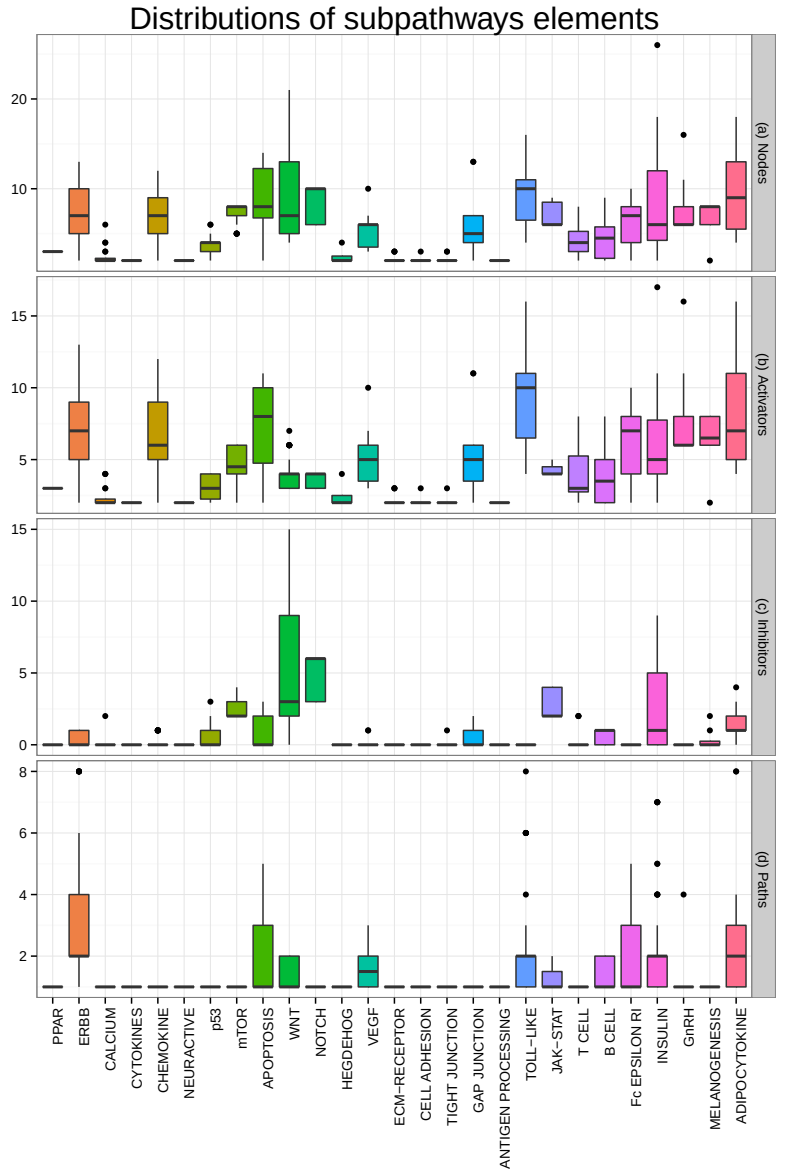


Figure 4.2: **Distribution of the different numbers of elements comprising a subpathway**. **(a)** The number of nodes comprising the subpathway. **(b)** The number of nodes acting as an activator in the subpathway. **(c)** The number of nodes acting as an inhibitor in the subpathway. **(d)** The number of paths in the subpathway

4.2 The calculation of probe set mixtures

Table 4.2 contains information about the specific number of arrays for each platform.

Platform	Specie	GEOid	Total available arrays
HGU133Plus2	*Homo Sapiens*	GPL570	3,034
HGU133A	*Homo Sapiens*	GPL96	5,293
Mouse1st	*Mus musculus*	GPL6246	683
MG430_2	*Mus musculus*	GPL1261	19,558
HGU133A_2	*Homo Sapiens*	GPL571	7,134
RG230_2	*Rattus norvegicus*	GPL1355	8,211

Table 4.2: **Available GEO arrays for each selected platform** The number of arrays avaliable in the Gene Expression Omnibus (GEO) database for each of the selected platforms.

The number of modeled platforms and the distribution selected to model each platform is represented in figure 4.3. Mixtures of gamma distributions gave better fits most for the probe sets in all platforms, and log-normal distributions were never selected for the mixtures (and this option will also be discarted for future analyses). The number of probe sets modeled for each platform was variable depending on how represented pathway's genes were in the platform.

As an example of the algorithm used to fit the mixture of distributions that best represents the behavior of the probe set, let us consider the probe set (*201655_s_at*). The density function for this probe set is represented in figure 4.4. Two different distributions can be observed in the density of this probe set, which are associated with an inactive and an active state respectively.

Once the developed algorithm is applied, two different fittings are considered, one associated with gamma distributions and the other associated with normal distributions. A good-of-fitness p-value associated with each of these fittings was also considered in order to choose the best fitted mixture of distributions. In the case of *201655_s_at* probe set, the fitted mixtures are: (Figure 4.5)

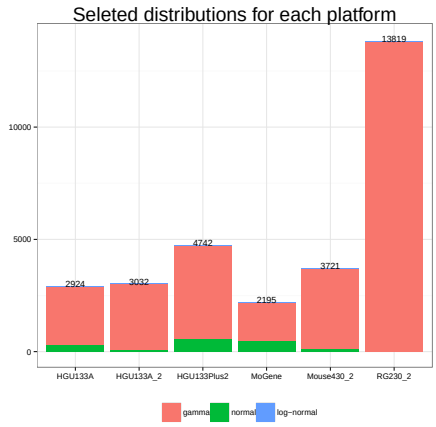


Figure 4.3: **The number of selected distributions for each platform.** The number of mixtures of gamma distributions is represented in red and number of mixtures of normal distributions is shown in green. There were no mixtures of log-normal distributions selected for any platform. The number over each bar represents the total number of probe sets modeled for each platform.

$$
\begin{aligned}
F(x) &= 0.393 \times \mathrm{Ga}(x|k = 59.401, \theta = 0.113) + \\
&\quad +0.607 \times \mathrm{Ga}(x|k = 132.299, \theta = 0.067) \\
F(x) &= 0.327 \times \mathrm{N}(x|\mu = 6.495, \sigma = 0.736) + \\
&\quad +0.673 \times \mathrm{N}(x|\mu = 8.793, \sigma = 0.843)
\end{aligned}
$$

In this case the normal distribution is selected since its good-of-fitness p-value is lower (p=0.16). Therefore, for future calculations of the probability of this probeset being active, the mixture of normal distributions with the estimated parameters will be used.

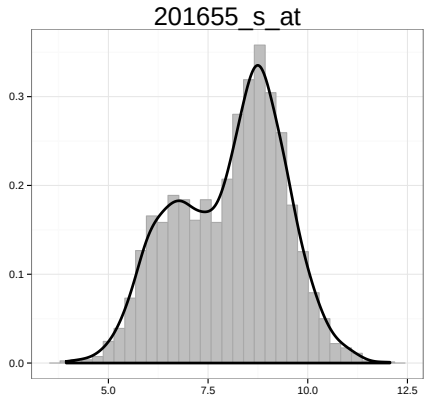
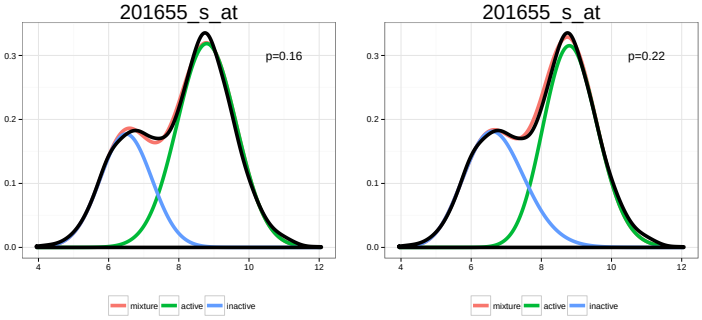


Figure 4.4: **Density function of the expression level** of the *201655_s_at* probe set



(a) Normal distributions (b) Gamma distributions

Figure 4.5: **Mixtures associated with the *201655_s_at* probe set. (a)**: Gamma distributions **(b)**: Normal distributions

4.3 Performance of the methodology

To test the proportion of false positives produced by any of the methods considered (multilevel modeling, t-test and Wilcoxon-test) we first created a large dataset containing samples in the same condition. It is important that the methodology developed must not find any significant subpathway in any of the combinations of samples chosen, therefore subpathways which were found to be significantly different between the same conditions were considered to be false positives. We use a large pediatric acute myeloid leukemia (AML) [110, 111] dataset which contains expression microarray data for 237 bone marrow and peripheral blood samples, hybridized to Affymetrix U133 Plus 2.0 arrays. This dataset was downloaded from the GEO database (identifier number *GSE17855*). N and M samples representing the control and disease samples respectively, were selected randomly from the AML dataset, and we performed, 500 simulations for each combination of sample sizes (see Table 4.3 and Figure 4.6).

N	M
5	5
5	20
10	10
10	100
50	50
50	100
100	100

Table 4.3: **Sample sizes** considered for the acute myeloid leukemia (AML) dataset simulations

As shown in Figure 4.6, multilevel testing gave rise to a huge proportion of false positives (significant subpathways) in all the cases considered, and therefore shows a low specificity and was not suitable for finding any real changes in the data. However, both the Wilcoxon and student-t tests behave similarly and had a level of false positives (type-I error) of less than 0.05, and consequently these two test have high specificity.

The proportion of false positives was also tested using simulated data. Specifically, we simulated uniform random values between zero and one representing the probabilities of the probe sets. Following this, normally distributed

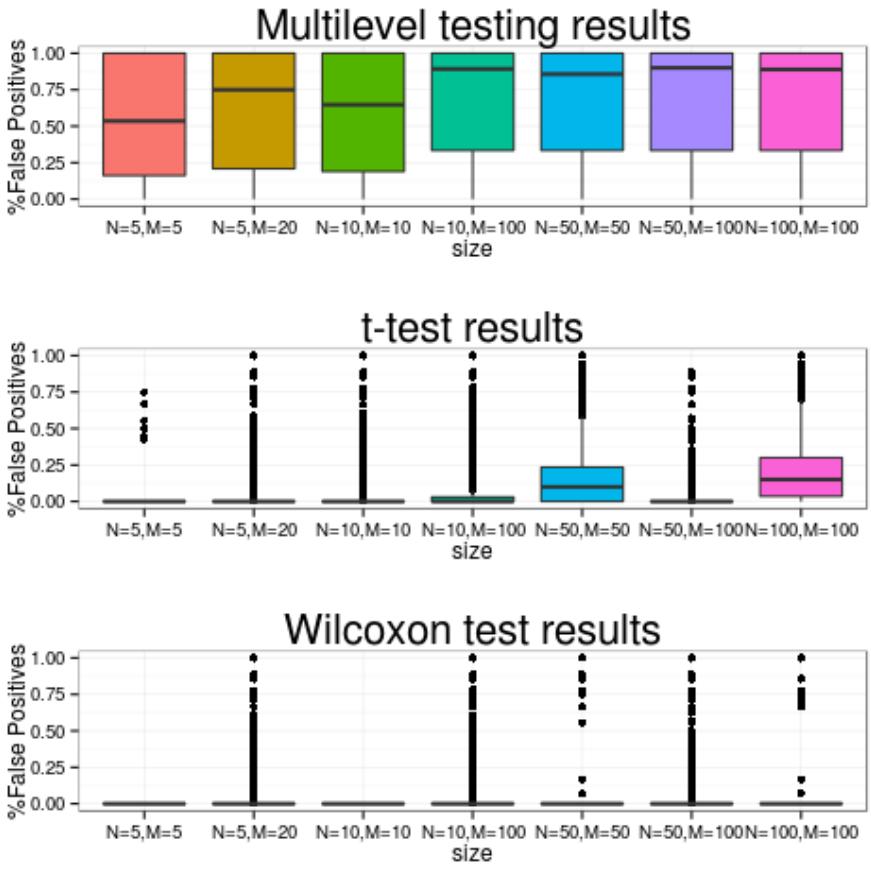


Figure 4.6: **Simulation results using the acute myeloid leukemia (AML) dataset** for the three different methods considered: multilevel modeling, the t-test and the Wilcoxon test

noise was added to the probability for each probe set probability. Different levels of noise are considered by using a normal distribution centered on zero and with different standard deviations ($\sigma = 0.25, 0.1, 0.05, 0.025, 0.01$). There were no big differences between different sample sizes, as shown in Figure 4.6,

and so the same sample size was used in these simulations. the proportion of false positives obtained with each method is shown in figure 4.7.

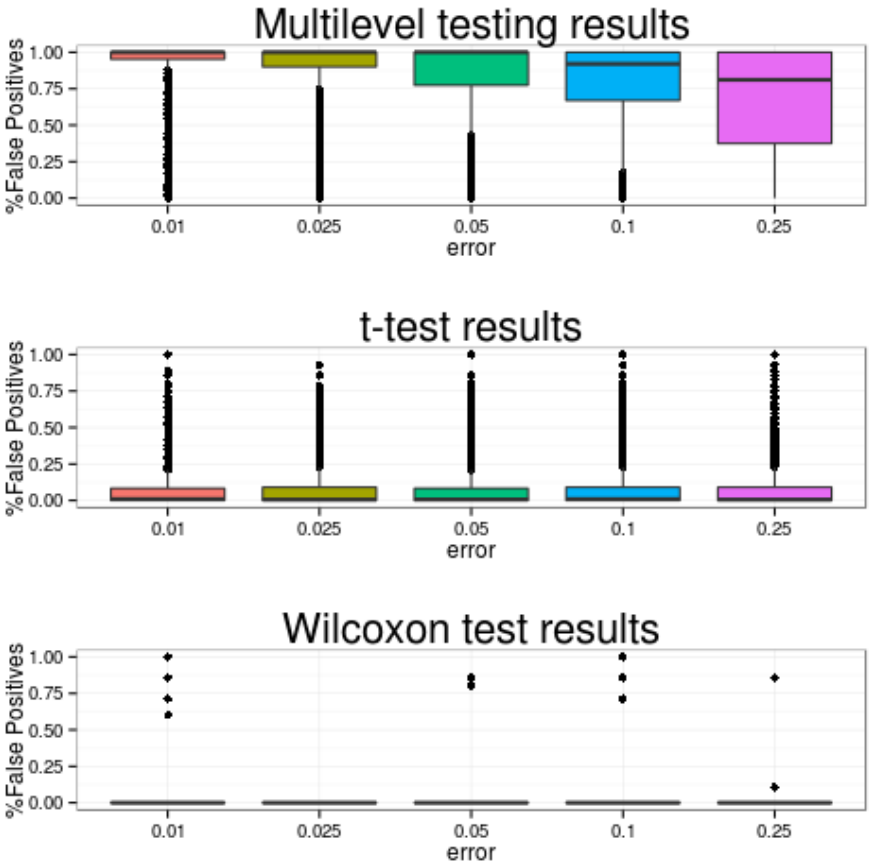


Figure 4.7: **Simulation results using simulated dataset** for the three considered methods: multilevel modeling, t-test and Wilcoxon test

Additionally, two other microarray datasets were used to check the predictive performance of the subpathway probabilities estimated using the approach we developed. This predictive performance can be considered represen-

tative of a low number of failures to detect real activations (false negatives)[7, 112]. We analyzed a breast cancer[113] study as well as and another dataset with increased expression in AML[114] (downloaded from the GEO database with identifiers *GSE27562* and *GSE9476* respectively), using PATHiPRED webtool[1][115]. Support Vector Machine (SVM)[116] was used to classify the samples and the accuracy of the classification was evaluated using ten-fold cross validation[117] taking the following parameters into account: proportion of correct classification (PCC) and area under the curve (AUC). The results obtained for both datasets are shown in table 4.4. Importantly, the accuracy of the predictors based on subpathway probabilities is close to one, thus showing that they have a high prediction power and consequently high sensitivity.

Dataset	PCC	AUC
Breast cancer	0.99	0.99
AML	0.96	0.95

Table 4.4: **Prediction results for the breast cancer and AML dataset.** Two parameters were considered to evaluate the accuracy of the prediction: the proportion of correct classification (PCC) and the area under the curve (AUC)

4.4 Results using real data

4.4.1 Colorectal cancer experiment

The method developed in this thesis was tested using a colorectal cancer (CRC) dataset [118] (GEO database identifier *GSE4107*) by comparing it with other methods (section 4.5). Some of the relevant results are highlighted in this section and a complete set of results from this experiment are included in Appendix B, although only some relevant results that probe the value of the proposed method are highlighted in this section.

CRC dataset was initially used to extract a relevant gene subset to early onset colorectal cancer. The sample contains 22 RNA samples extracted from the colonic mucosa of healthy controls (10 samples) and patients(12 samples)who were aged 50 or less and were Chinese. The RNA was analyzed using the *Affymetrix Human Genome U133 Plus 2.0 Array.*

[1]See 5.1 for a complete explanation of this tool

Several pathways are known to be evolved in the development of colorectal cancer, and we will specifically focus on three of these in this section: *VEGF signaling pathway*, *JAK-STAT signaling pathway* and *WNT signaling pathway*.

It is known that, human colorectal tumors produce vascular endothelial growth factor (VEGF) whose expression is up-regulated in tumor cells by both cyclooxygenase-2 (COX-2) and PGE2 and is directly correlated to neoangiogenesis and clinical outcome[119]. The COX-2 gene is included in node the named PTGS2. In addition, the increase of the activity of COX-2 is connected with PGI2 production in colorectal carcinomas[120]. According to our results subpathway that ends in this node is significantly more activated in colorectal cancer patients (See figure B.13). VEGF-A also increases the vascular permeability to plasma and plasma proteins, that is a characteristic property of tumor microvasculature, a critical early step in tumor stroma generation[121]. In line with this, as shown in Figure B.13, a subpathway whose endingpoint function is permeability is significantly more activated in colorectal cancer patients.

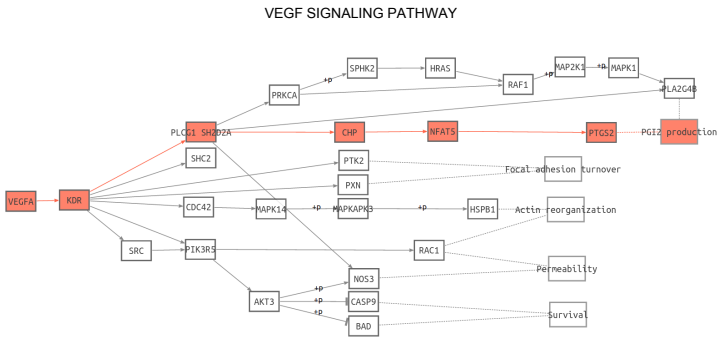


Figure 4.8: Results from the ***VEGF signaling pathway*** in the colorectal cancer study

The *JAK-STAT signaling pathway* is also affected in colorectal cancer, its inhibition reduces tumor cell invasion in colorectal cancer cells[122, 123]. Significant subpathways identified using the methodology developed in this thesis include functions which are closely related with cancer development such as growth, proliferation, fate determination, development, immunity, cell cycle and antiapoptosis (Figure 4.9).

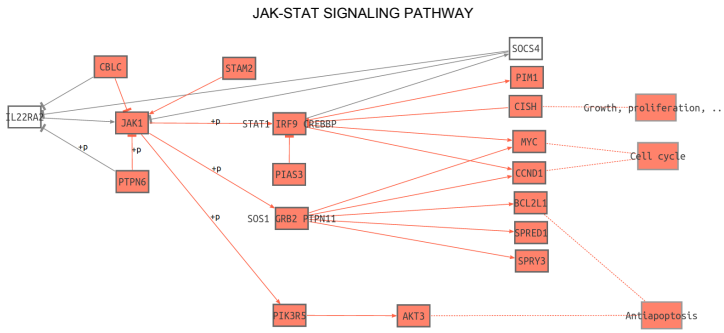


Figure 4.9: results from the ***JAK-STAT signaling pathway*** in the colorectal cancer study

Finally, the *WNT signaling pathway* (Figure4.10) is frequently affected in colorectal cancer. This pathway comprises by three different pathways: canonical (*Wnt/β-catenin cascade*), and non-canonical (*Wnt/Planar cell polarity(PCP)*, and *Wnt/Ca2+ mediated*). The results obtained after applying our method revealed significantly increased non-canonical pathways activity. Genes belonging to the *Wnt/PCP pathway*, such as RhOA, RAC and JNK are known to be upregulated in CRC cancer [124], which explains the increase in the activity of the subpathway leading to JNK[2].

[2]labeled as MAPK8 in the pathway representation

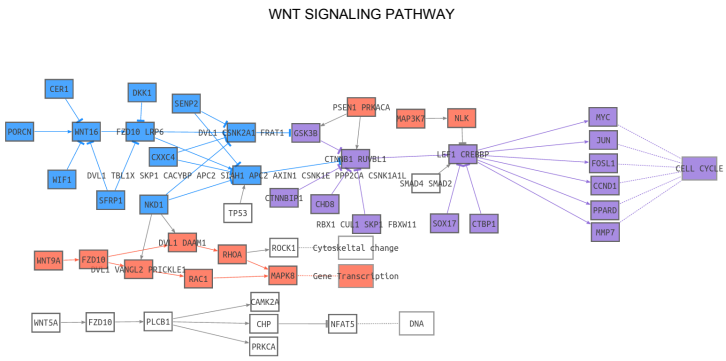


Figure 4.10: Results from the ***WNT signaling pathway*** in the colorectal cancer study

4.4.2 Models of obesity

In a previous study adipose tissue macrophages (ATMs) obtained at 5 ans 16 weeks of age from wild type (WT) and ob/ob mice were characterized[125]. ob/ob is a mutant mouse model that eats excessively and becomes obese and is commonly used in type II diabetes studies. We used this data to study this ATM characterization from a subpathway-based point of view, and obtained compatible results. The dataset used for this analysis was downloaded from the GEO database under identifier *GSE36669*. This dataset included 15 samples describing 2 genotypes (WT and ob/ob) at 2 time points (5 and 16 weeks) (as shown in Table 4.5). The RNA samples were analyzed using *Affymetrix Mouse Gene 1.0 ST Array* microarray.

genotype	time point	#samples
WT	5 weeks	4
	16 weeks	4
ob/ob	5 weeks	3
	16 weeks	4

Table 4.5: **Experimental design of ATM dataset.** 15 samples describing 2 genotypes (WT and ob/ob) and 2 time points (5 and 16 weeks).

A predominant M2 anti-inflammatory phenotype was observed in 16-weeks-old WT ATMs as well as 5-weeks-old ob/ob ATMs. Conversely, 16 weeks old ob/ob ATMs had switched to a predominantly M1 proinflammatory phenotype, which is associated with severe insulin resistance, diabetes, and proinflammatory macrophages in adipose tissue. In contrast, after 16 weeks, the WT ATMs could still control their carbohydrate metabolism and progressively expanded their adipocyte tissue. This process is represented in the comparison between 5-weeks-old and 16-weeks-old wild type mice. Concretely, *WNT signaling pathway* (Figure 4.11) significantly activated the *Wnt/β-catenin* canonical which triggered the cell cycle. The role of canonical pathway in tissue remodeling by weight gain has already been identified [126, 127].

The *VEGF signaling pathway* also contributes to remodeling adipose tissue during age-related growth expansion[128] (Figure 4.12). Its activity increases as demand for adipose tissue expansion increases as well as in hypoxia which produces an increase in the adipose tissue vascularization.

Finallly, subpathways in the *Apoptosis* pathway are also activated in 16-

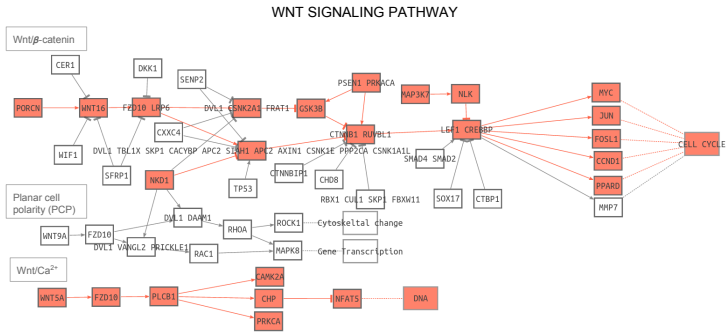


Figure 4.11: **Results from *WNT signaling pathway* in the adipose tissue macrophages (ATM) dataset.** Wild type 5 weeks old were compared with wild type 16 weeks old mice and we found a significant activation of subpathways which trigger cell cycle in the canonical *Wnt/β-catenin* pathway.

week-old ob/ob mice compared to the Wild type animals (Figure 4.13). Specifically, caspase substrate cleavage is activated, as well as subpathways related with survival, apoptosis and degradation. These results correlate with the huge amount of fat depositio, inflammatory responses, and adipocyte crowns of apoptotic adipocytes which are characteristic of the ob/ob mouse phenotype as it ages.

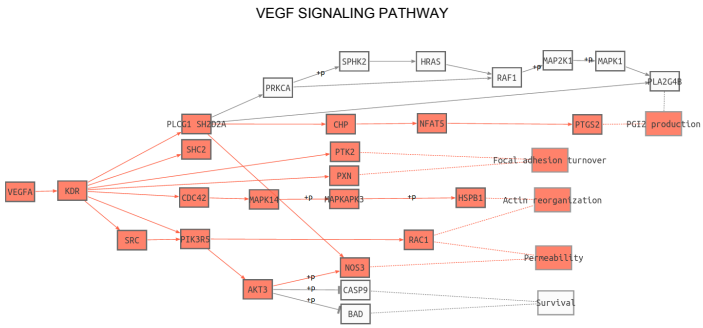


Figure 4.12: **Results from *VEGF signaling pathway* results in the adipose tissue macrophages (ATM) dataset.** Wild type 5 weeks old were compared with wild type 16 weeks old mice and we found a significant activation of subpathways ending on PGI2 production, Focal adhesion turnover, actin reorganization, and permeability

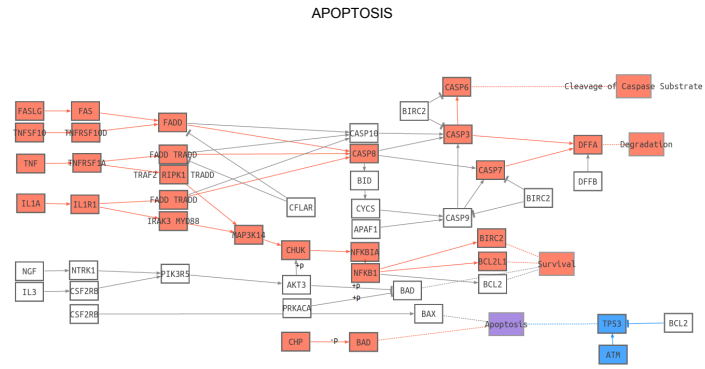


Figure 4.13: **Results from *Apoptosis* in the the adipose tissue macrophages (ATM) dataset.** Wild type 16-weeks-old were compared with ob/ob 16 week old mice and we found general activation of several subpatwhays in *Apoptosis* pathway which trigger opposite functionalities as apoptosis or survival.

4.4.3 Stem cells

This dataset was used in [129] to characterize changes in human hematopoietic stem cells (HSCs) and progenitor cell populations during aging. It was downloaded from GEO database under the identifier *GSE32719* and includes 27 HSC samples from healthy, hematologically normal (14 samples), young (5 samples), and elderly (8 samples) human bone marrow. The analysis carried out in the original study suggested that there is a coordinate deregulation in genes associated with cell cycle, hematological system development and function, cellular growth and proliferation, and humoral inmune response in elderly HSC compared with younger ones[130]. This behavior was highlighted when we analyzed the data using our method to show that a subpathway triggering proliferation, differentiation, and Immune response is significantly activated in older HSCs compared with young HSCs (Figure 4.14). Therefore, the original study discovered gene expression signatures and our developed method uncovered the activation mechanisms used by the cell.

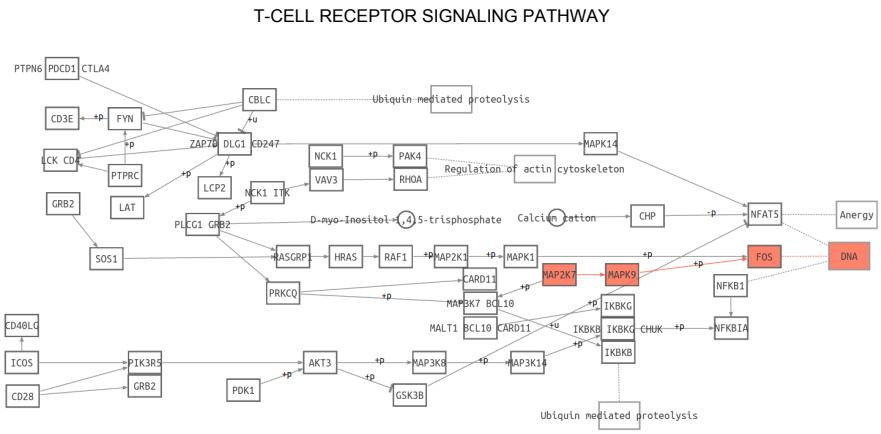


Figure 4.14: Results from ***T-cell receptor signaling pathway*** **in the human hematopoietic stem cells (HSC) dataset.** Subpathway which trigger Proliferation, Differentiation and Immune response is significantly activated in older HSC compared with young HSC.

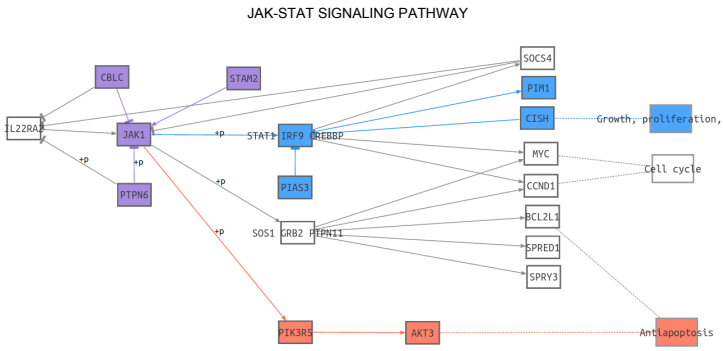
4.4.4 Fanconi Anemia - Cross-talk between pathways

Similarly to genes, signaling pathways do not act alone, but rather they are connected among themselves to form a global pathway that controls cell signaling processes. Therefore, results in one pathway can potentially reveal the precise mechanisms which trigger a particular biological response in another pathway, that is, the functional consequences in one pathway can provoke the activation/inhibition of subpathways in another pathway[3]. Observing the combined behavior of signaling pathways in a diseased cell can eventually reveal interesting details about the overall mechanism of the disease and thus help in the inference of ways to intervene.

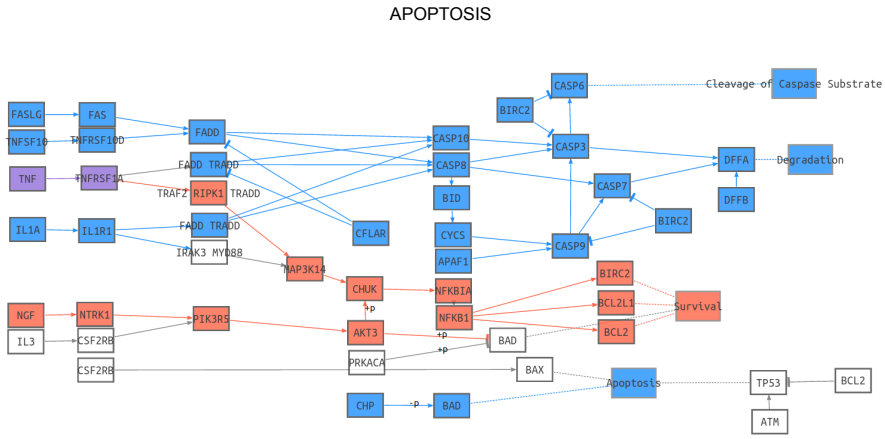
To illustrate this concept, we analyzed a dataset from Fanconi Anemia (FA), a disease in which signaling plays an important role. This dataset was used in [131] to identify transcriptional differences between bone marrow cells from 11 normal volunteers and from 21 children and adults with Fanconi anemia. It was downloaded from the GEO database under the identifier *GSE16334*. FA is a rare chromosome instability syndrome that is characterized by aplastic anemia as well as a susceptibility to cancer and leukemia [132]. Disruption of apoptotic control is a hallmark of FA, ans it has been proposed that this explains the phenotype of the disease to some extent[133] and one of the genes related to the disease, FANCC, is thought to be involved in Jak-STAT signaling and apoptotic signaling[132]. Modeling of both pathways using the data and our method allows us to pinpoint the mechanisms by which *Jak-STAT signaling pathway* specifically triggers one of the survival circuits in the *apoptosis pathway* that eventually causes the disease (Figure 4.15).

Centering on the Jak-STAT results (Figure 4.15a), the subpathways leading to Growth Proliferation (via the PIM1 and CISH genes) are inhibited in FA patients and the subpathways leading to Antiapoptosis (via the AKT3 gene) are activated. This antiapoptotic behavior is directly associated with the *Apoptosis pathway* (Figure 4.15b). Three functionalities appear to be significantly inactive in FA: Cleavage of caspase substrate, Degradation, and Apoptosis (via the BAD, DFFA and CASP6 genes respectively); whereas Survival (via BCL2L1, BIRC2 and BCL2 genes) is significantly active. Therefore, the final consequence is that antiapoptotic pathway activity in FA is higher than in normal cells, although, apoptosis does still occur[133], most likely through the circuits that ending at the BAX and/or TP53 genes(Figure 4.15b), whose activity is not significantly different between normal and FA patients. This observation suggests that known the features of FA, such as hypersensitivity

[3]For example, the final gene of one subpathway can be initial signal in another pathway



(a) JAK-STAT signaling



(b) Apoptosis

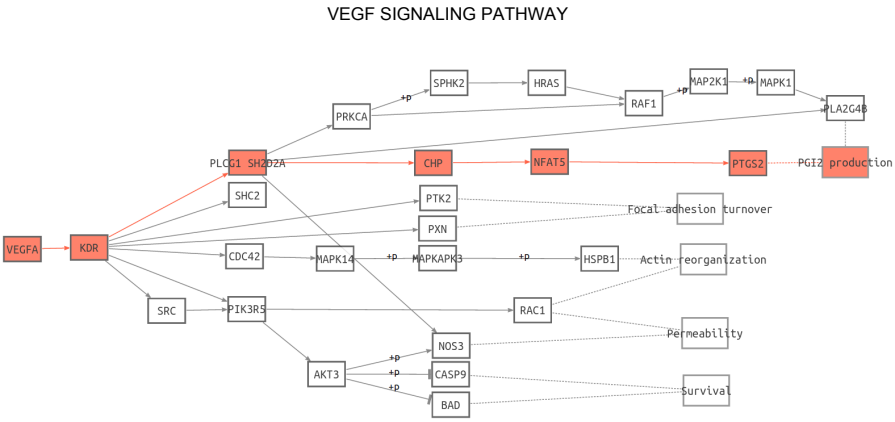
Figure 4.15: Models of the **(a) JAK-STAT** and **(b) Apoptosis** pathways in FA along with the corresponding significant changes in the signaling circuit activities.

to DNA cross-linking agents [134, 135] or chromosomal instability [135] may be a consequence of the abnormal survival of cells with damaged DNA. In fact, recent reports have confirmed that FA proteins participate directly in canonical signaling pathways that influence survival and hematopoietic cells self-replication[131].

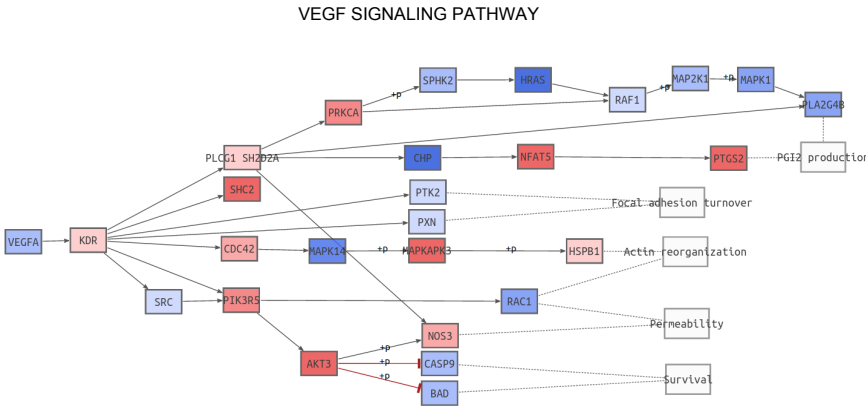
4.5 Comparison with other approaches

The definition of subpathways proposed in this thesis is not exactly the same as any of the other pathway analysis approaches, and so it is not possible to compare results in a straightforward way. However, in this section, the results obtained from analyzing the CRC dataset (Section 4.4.1) were used to compare the approach described here with other methods that are typically used to analyze expression data in pathway context.

The simplest analysis that can be done in a pathway context is the graphical comparison of the differential behavior of each individual gene in a pathway context. We calculated differential expression changes using limma[100] R-package and represented them in a pathway-context (Figure 4.16b) using the following scale: red for up-regulated, blue for down-regulated and white for no differential expression, where intensity of the color depended on the significance of the differential expression. The differential expression and *PATHiWAYS* results for the *VEGF signaling pathway* are displayed in Figure 4.16. It is worth noting that it is not possible to extract conclusions about subpathway activities from the information obtained. It is common that different gene up- and down-regulations compensate each other so that no final change is produced in any of the subpathway activities (Figure 4.16a). Therefore, significantly activated or deactivated genes with no direct effect in this pathway, may be acting in another pathway where its differential activity is not compensated by the other genes acting in the network.



(a) subpathway-centered



(b) gene-centered

Figure 4.16: **Methods comparison.** Comparison of the results obtained applying: **(a) subpathway-centered approach**, *PATHiWAYS* results for the *VEGF signaling pathway* and **(b) gene-centered approach**, representation of the differential expression of each gene where red represents over-expression and blue under-expression. The intensity of the color is related with the magnitude of the differential expression

As described in section 1.5, the methods developed for pathway analysis can be divided into four generations. Methods included in first three generations provide an enrichment analysis of the pathway taken as a whole, without distinguishing between the different subpathways that compose it,meaning that our results are not directly comparable with the results obtained with these methods. However, we used a method that corresponds with each of these generations to analyze the CRC dataset and the results obtained were compared with the PATHiWAYS results as far as possible (Table 4.6).

To compare the PATHiWAYS results with ORA tests (first generation methods) we use the FatiGO[44] web-tool included in the Babelomics[136] platform. The input for the tool is a list of genes ranked by the statistics obtained from a differential expression experiment. The upper and lower part of this list (corresponding with the most up- and down-regulated genes) can be compared to each other or with the complete list of genes in the genome. This tool returns a list which is enriched in the genes which map to each pathway. The results obtained are included on Table 4.6 in the columns named: *UP vs DOWN* which compares upregulated and downregulated genes; *UP vs ALL* which compares upregulated genes with all the genes included on the microarray[4]; and *DOWN vs ALL* which compares downregulated genes with all the genes included on the microarray.

Results from a functional class scoring (FSC) test, representative of the family of GSEA tests which does not require gene pre-selection were compared with PATHiWAYS results. More specifically, the Fatiscan[137] web-tool included in Babelomics platform was used to asses which pathways are enriched in genes with decreased gene expression. The results obtained with this method are included in table 4.6 in the *GSEA* column.

Finally, to compare the results from our method with those from third generation of pathway analysis methods, the Signaling Pathway Impact Analysis (SPIA)[83] approach was selected. This method uses the information from a list of differentially expressed genes and their log-fold changes together with signaling pathways topology, in order to identify the pathways which are most relevant to the condition being studied. The CRC data was analyzed using SPIA R-Package[84] and results obtained are included in Table 4.6 under the *SPIA* column.

First column in Table 4.6 contains the name of the pathway. The next three columns, collectively labeled as circuits, list the number of sub-pathways

[4]We use a personal annotation to avoid noise which include only genes measured by the microarray

in the pathways (Total) and the number of them significantly activated in cases with respect to controls (Case) or vice versa (Control) in the comparison, respectively. The three next columns, collectively labeled as SEA, list the results of a conventional functional enrichment test in three situations: UP vs DOWN) when the significantly upregulated genes are compared to the significantly downregulated genes, UP vs ALL) when significantly upregulated genes are compared to rest of genes, and DOWN vs ALL) when significantly downregulated genes are compared to the rest of genes. UP, DOWN and ALL means where the major part of the pathway lies in the comparison. Significantly up- and downregulated genes are obtained by a conventional t-test with multiple test adjustment as implemented in the Babelomics program. Although the trends of the results are coincident with the other analyses, none of them resulted significant. The column labeled GSEA contains a version of GSEA test implemented in the Babelomics program. The * and the boldface indicate the trend is significant according to the test. The last column, labeled as SPIA, contains the result of the application of the pathway impact analysis

From table 4.6 we can conclude that no discrepancies between the results obtained using the SPIA (that take pathway structure into account) and the PATHiWAYS method stood out. In some cases, SPIA's behavior is closer to taht of GSEA, probably because that both methods return a global score for the pathway. However, the method proposed in this thesis goes further to detect specific aspects of pathway activity. For example the *VEGF signaling pathway* is known to be active in cancer, and its inhibition has been suggested as an anticancer therapy [138]. Several subpathways in this pathway were detected as significantly active by our approach but not by the alternatives[5]. Similary, the *Jak-STAT signaling pathway* is implicated in CRC, and its disruption reduces tumor cell invasion[122]. In this case, both GSEA and our method[6] detected this behavior, but SPIA did not. Therefore, the PATHiWAYS method detects pathway activity detected by other methods, but also detects the activation of specific signaling functionalities that other methods are unable to detect.

[5] Figure B.13
[6] Figure 4.9

	PATHiWAYS			SEA				
PATHWAY	**Total**	**Case**	**Control**	**U vs D**	**U vs A**	**D vs A**	**GSEA**	SPIA
PPAR	106	3	19	DOWN	ALL	ALL	DOWN*	INH*
ERBB	139	11	2	UP	UP	DOWN	DOWN*	INH
CALCIUM	20	2	2	DOWN	UP	DOWN	UP*	ACT
NEUROACTIVE	7	0	0	UP	UP	ALL	UP*	ACT
APOPTOSIS	28	0	0	DOWN	ALL	DOWN	DOWN*	INH
WNT	37	6	6	UP	UP	DOWN	DOWN*	INH
NOTCH	14	0	0	DOWN	ALL	ALL	DOWN	INH
VEGF	10	2	0	DOWN	ALL	ALL	DOWN	INH
CELL ADHESION	43	6	3	UP	UP	ALL	UP*	–
GAP JUNCTION	17	4	0	UP	UP	ALL	UP*	ACT
ANTIGEN PROCESSING	6	0	0	DOWN	ALL	DOWN	UP	ACT
TOLL-LIKE	103	0	0	UP	UP	DOWN	UP*	INH
JAK-STAT	7	7	0	DOWN	ALL	DOWN	UP*	INH
B CELL	10	0	0	UP	UP	ALL	UP*	INH
Fc EPSILON RI	7	0	0	UP	ALL	ALL	DOWN*	INH
INSULIN	54	1	0	DOWN	UP	ALL	DOWN*	INH
GnRH	9	0	0	UP	UP	ALL	DOWN*	ACT
MELANOGENESIS	8	1	0	UP	ALL	ALL	UP	ACT
ADIPOCYTOKINE	31	0	2	DOWN	ALL	ALL	UP	INH

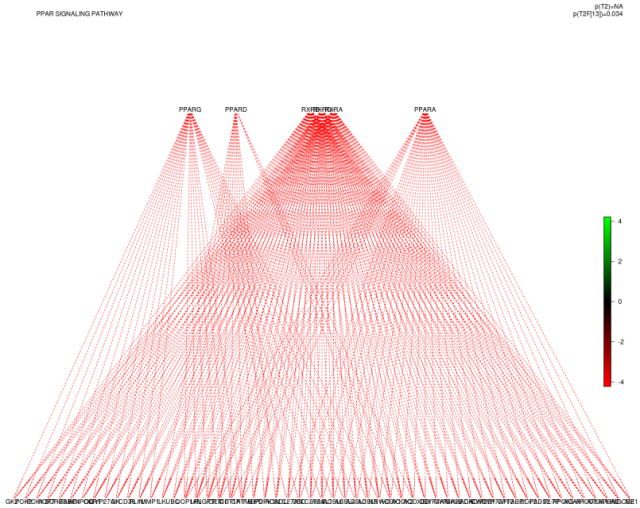
Table 4.6: **Comparison of results obtained with PATHiWAYS method with other approaches.**

Finally, two subpathway-based methods were run with the CRC dataset and compared with our results: `DEGraph`[33] and `clipper`[34], concretely their web-tool called `GraphiteWeb`[35]. However, these methods also have a different definition of subpathway, making the results no directly comparable. Concretely, both methods define a subpathway as a clique (Section 1.5).

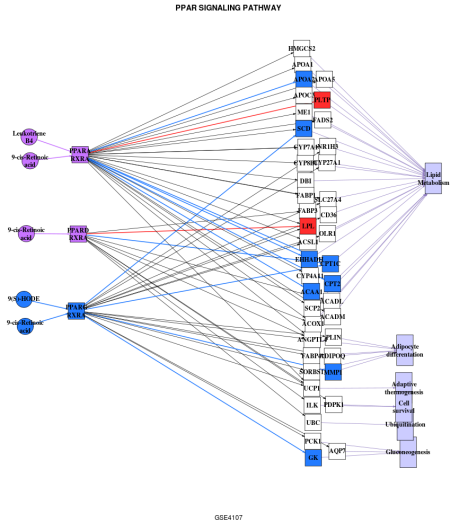
Two examples of DEGraph results are shown in figures 4.17,4.18.

The Figure 4.17shows the significant DEGraph and PATHiWAYS results. DEGraph returns the whole pathway as a significantly affected module (Figure 4.17a), whereas with our method highlights different behaviors in pathway depending on the different stimulus and responses (Figure 4.17b). As previously mentioned in section 3.1, this example shows incorrect KEGG information modeling, because DEGraph considers the PPAR and RXR proteins to be independent signal entries in the network, even though in reality they form a complex and behave cooperatively.

In the case of the *ErbB signaling pathway* (Figure 4.18), two different pathways are highlighted as significant by the DEGraph results (Figures 4.18a, 4.18b). The first subpathway represents the central part of the pathway and was also located by the PATHiWAYS approach (concretely subpathways ending in JUN and ELK1). The second subpathway identified as significant by DEGraph is in fact formed by two separate nodes which contain several genes. DEGraph considers them to be independent, but according to KEGG information ans as previously discussed in section 3.1, they are alternative signal transduction pathways.

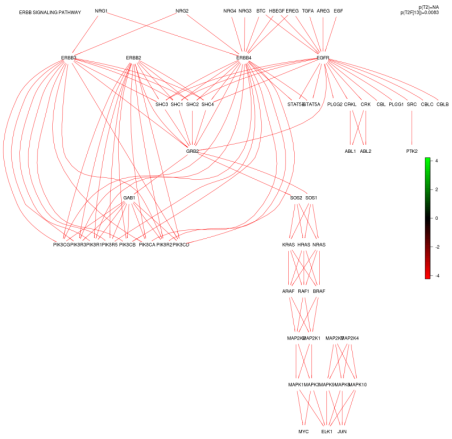


(a) DEgraph results

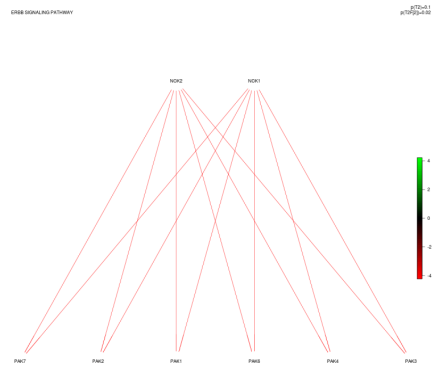


(b) PATHiWAYS results

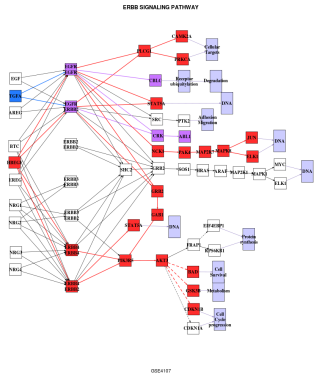
Figure 4.17: **DEGraph and PATHiWAYS significant results on *PPAR signaling pathway***



(a) First DEgraph significant subpathway



(b) Second DEgraph significant subpathway



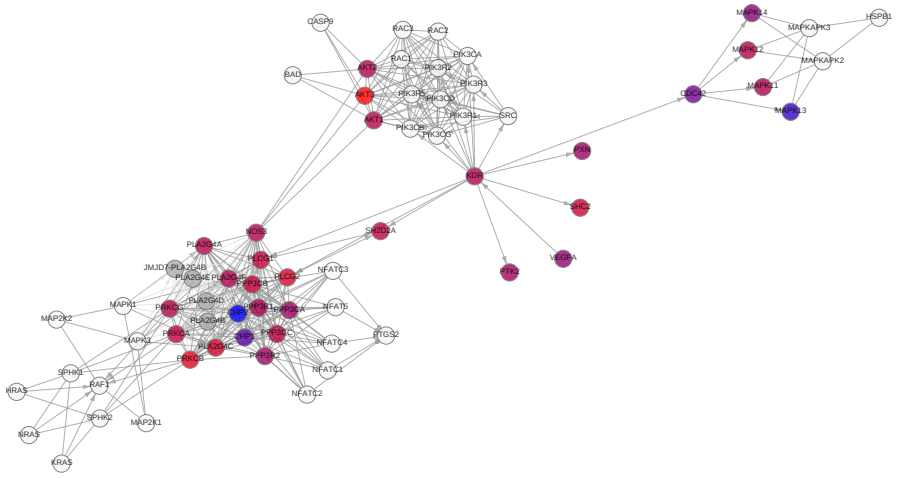
(c) PATHiWAYS results

Figure 4.18: **DEGraph and PATHiWAYS significant results from *ErbB signaling pathway***

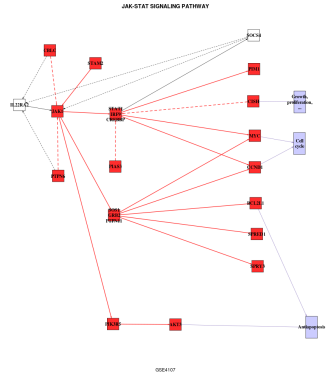
The significant results from GraphiteWeb are shown in Table B.19[7]. A large number of pathways, such as Cocaine addiction, African trypanosomiasis, Long-term depression, Salivary secretion, and many others that are completely unrelated to cancer are obviously false positives. Forexample, in the *VEGF signaling pathway* (Figure 4.19), clipper only detects a portion of the subpathway leading to the production of COX2 via PTGS2 (prostaglandin G/H synthase and cyclooxygenase), which is known to be activated in CRC [139, 119] and is detected by our approach.

Finally, it is important to highlight that in last months Li published a new paper comparing subpathway-based methods where PATHiWAYS were included[140]. He propose a new approach called sub-SPIA and compare it with some of the existing methodologies: SPIA, clipper, DEGraph and PATHiWAYS, and he concludes that pathways identified by sub-SPIA and PATHiWAYS generally play important roles in the entire system and reflect the hub characteristic of the identified pathways. But, again, subpathway definition lacks of biological interpretation on this approach.

[7] A complete version of the results as well as representations can be consuted in `http://graphiteweb.bio.unipd.it/results/f03dfb1161af34816216f5d4e043d9ccc4c66cfbfca60cdd8c490060b12266b1/`



(a) GraphiteWeb results



(b) PATHiWAYS results

Figure 4.19: ***VEGF signaling pathway*** **significant results from GraphiteWeb and PATHiWAYS approaches**

4.6 Drugs behavior - CAMDA challenge

CAMDA (Critical Assessment of Massive Data Analysis) is a well-recognized annual conference recently called the 'Olympics of Genomics'. This is a contest to introduce and evaluate new approaches and solutions to the 'Big Data' problem[8]. Therefore, CAMDA is a data analysis challenge which focuses on the analysis of big heterogeneous data sets. These data sets are provided by the organizers along with some questions about the data which the applicants must answer. Research groups in the field of bioinformatics, data analysis, and statistics propose new techniques for handling and processing these data sets, the combination of multiple data sources, and computational inference, and best applications are invited to present and discuss their methods during the conference.

We use the approach developed in this thesis to analyze the data provided for 12th annual conference (CAMDA 2013). The data provided in one of the challenges of this contest was the TGP dataset from the Japanese Toxicogenomics Project[141] that contains over $> 21,000$ arrays for rats treated with mainly human drugs and profiled using the *Affymetrix RAE230_2.0 GeneChip*. In this case, only the data for liver tissue (both in vivo and in vitro) is provided. The challenge questions for these data were:

Question 1 Can we replace the animal study with in vitro assay?

Question 2 Can we predict the liver injury in humans using toxicogenomics data from animals?

We tested first question using the PATHiWAYS[89, 88] approach. First, more than 10,000 samples from the *Affymetrix RAE230_2.0 GeneChip* were downloaded from the GEO[25] database in order to estimate the mixtures for each of the probe sets measured (as explained in Section 3.2). Rat pathways information was also downloaded from the KEGG[27] database and modeled as indicated in section 3.1.

We then carried out independent comparisons between all of the doses of each drug tried in vitro and in vivo, and studied which circuits in which pathways displayed a significant change in activity when the drug was induced, as well as the type of change (activation or inhibition).

A total of 931 different circuits from all the pathways were affected by one or more drugs. Cell lines were affected more by drugs than the corresponding

[8] Referring to the explosion of new techniques that generate biological datasets (See Section 1.3)

in vivo counterparts (by an average of more than 25%) as shown in Figure 4.20,the mean number of drug affected subpathways in vitro are represented in blue, in vivo in red, and the intersection between them n orange. This indicates that, despite the disparity in global gene expression, the global behaviors are, probably, not as radical as was indicated by Lukk[142].

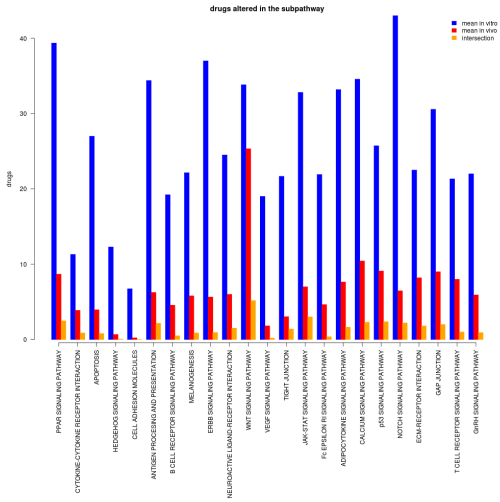


Figure 4.20: **Mean number of drug affected subpathways per pathway**, where the mean number of drug affected subpathways in vitro are represented in blue, in vivo in red ans the intersection between them in orange

However, only 207 (out of a total of 931) stimulus-response circuits, corresponding to almost all the pathways (20 out of a total of 23 modeled[9]) (represented in Table 4.7) sowed any coincident patterns of activation in response to the several drugs tried. Almost half of the drugs tried (58 out of 132) caused an identical effect both in vitro and in vivo in at least one circuit of at least one pathway.

Pathway Name	Drugs
PPAR signaling pathway	bendazac, benzbromarone, benziodarone, clofibrate, fenofibrate, furosemide, gemfibrozil, simvastatin, sulfasalazine, WY-14643
p53 signaling pathway	colchicine, disopyramide, ethionine, moxisylyte, nitrosodiethylamine, propylthiouracil, puromycin aminonucleoside, quinidine
Citokine-Citokine receptor interaction	diazepam
Apoptosis	hydroxyzine, nitrofurantoin
Hedgehog signaling pathway	
Cell adhesion molecules	caffeine, aproxen, nitrofurazone, tacrine, colchicine, gentamicin
Antigen processing and presentation	flutamide, puromycin aminonucleoside
B cell receptor signaling pathway	nimesulide, nitrofurazone, chloramphenicol, colchicine, mexiletine, gentamicin, hydroxyzine, sulpiride
Melanogenesis	doxorubicin, isoniazid
ERBB signaling pathway	hydroxyzine, nitrofurantoin, colchicine, ethionine, colchicine, caffeine
WNT signaling pathway	caffeine, ibuprofen

[9] See Table 4.1

VEGF signaling pathway	acetamidofluorene, cyclophosphamide, danazol, diazepam, ethambutol, ethinylestradiol, ibuprofen, cyclosporine A, diazepam, ajmaline, ethinylestradiol, ethambutol, nitrofurantoin, nitrofurantoin
Tight Junction	caffeine, cisplatin, naproxen, sulindac, ethionine, gentamicin, monocrotaline, puromycin aminonucleoside
Jak-STAT signaling pathway	diclofenac, disopyramide, furosemide, ibuprofen, sulindac
Fc epsilon RI signaling pathway	colchicine, ethionine, gentamicin, penicillamine, valproic acid
Adipocytokine signaling pathway	diclofenac, naphthyl isothiocyanate, naproxen, colchicine
Calcium signaling pathway	ethionine, hydroxyzine, caffeine
Notch signaling pathway	methimazole, naproxen
ECM-receptor interaction	nifedipine
Gap junction	carbon tetrachloride
T cell receptor signaling pathway	colchicine, ethionine, gentamicin, penicillamine, valproic acid, caffeine, disopyramide, naproxen, sulindac, naphthyl isothiocyanate, hydroxyzine, coumarin
GnRH signaling pathway	disopyramide, naproxen, iproniazid

Table 4.7: **CAMDA results**. Table showing the pathways with coincident patterns of activation between in vitro and in vivo samples, in response to several of the drugs tried.

The results in the context of the *PPAR signaling pathway* are represented in Figure 4.21. The subpathway activated by: benzbromarone, clofibrate, fenofibrate, and WY-14643 is shown in orange; subpathway activated by: benzio-

darone, and sulfasalazine is shown in purple; andsubpathway activated by: bendazac, benzbromarone, fenofibrate, gemfibrozil, simvastatin, and WY-14643 is shown in pink. The orange and purple subpathways produced the same functional consequence: *Lipid metabolism* activation, and the pink subpathway provoked *Adipocyte differentiation*.

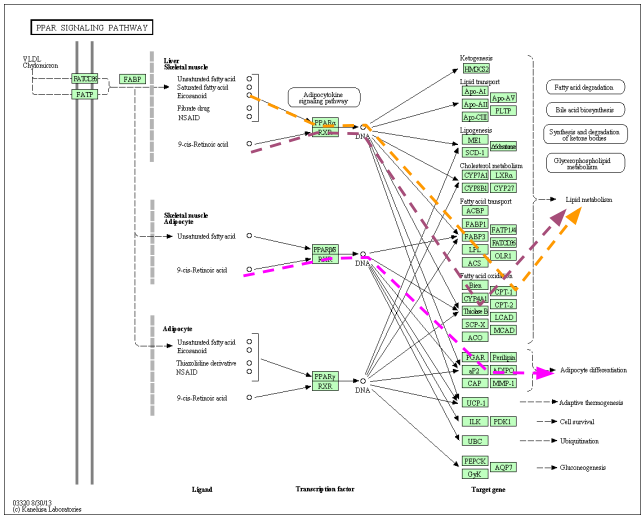


Figure 4.21: **Drugs showing the same behavior between the in vitro and in vivo samples as represented in a pathway context.** The subpathway activated by: benzbromarone, clofibrate, fenofibrate, and WY-14643 is shown in orange; subpathway activated by: benziodarone, and sulfasalazine is shown in purple; andsubpathway activated by: bendazac, benzbromarone, fenofibrate, gemfibrozil, simvastatin, and WY-14643 is shown in pink. The orange and purple subpathways produced the same functional consequence: *Lipid metabolism* activation, and the pink subpathway provoked *Adipocyte differentiation*

Using pathways to assess drug responses has a number of limitations. Firstly, there are drugs (half of the drugs tested here) that do not affect the set of modeled signaling pathways and therefore their effects remain undetectable. Secondly, in other cases, strong responses, which are mainly observed in vitro, mask the induction or repression of common circuits that might be useful to predict drug activity.

However, despite these limitations, our results suggest that pathway models can offer an interesting alternative to other "black box" methods for drug activity prediction. More detailed modeling of cell activity, including metabolic pathways and other aspects such as regulation, protein interaction, etc., will probably increase their predictive accuracy while also providing valuable information on the mechanisms of drug action.

4.7 The *PATHiWAYS* tool

PATHiWAYS[89] web-tool (See figure 4.22) was developed in order to carry out the special analysis approach developed in this thesis in the simplest and most user-friendly way possible.

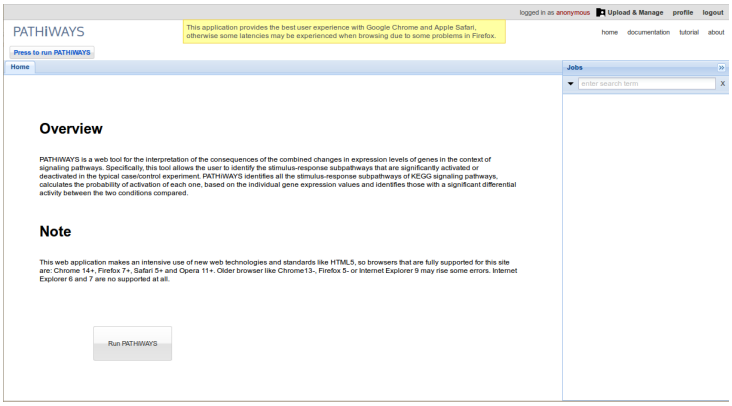


Figure 4.22: **Main page of *PATHiWAYS* web-tool**

PATHiWAYS is a web-based signaling pathway analysis and visualization system that infers changes in signaling that affect cell functionality from gene expression values obtained in typical expression microarray case-control experiments. This tool allows to the user to easily analyze microarray expression data from five different platforms, three for human and two for mouse (the platforms available are shown in Table 4.8).

The user can use expression microarray raw data (.CEL files) or a rma normalized expression matrix as input data. The analysis method developed in this thesis is applied to these data given as experimental condition file indicating which sample belongs to each condition. The user can choose the

Specie	Platform
Homo sapiens	Affymetrix Human Genome U133 Plus 2.0 Array Affymetrix Human Genome U133A Array Affymetrix Human Genome U133A 2.0 Array
Mus Musculus	Affymetrix Mouse Gene 1.0 ST Array Affymetrix Mouse Genome 430 2.0 Array

Table 4.8: **Platforms** which were modeled in *PATHiWAYS* tool

function used to summarize the probability of multiple genes being in a node (mean, median, maximum, minimum and 90th, 95th and 99th percentiles), the test applied (multilevel or Wilcoxon) and the pathways used to do the analysis (from the ones modeled for each species).

The output is divided into three sections. A summary of the input data for the analysis is shown in the first one. The second one is a clickable index of the selected pathways and the item summary. Here, the user can choose the results will be shown in the last section. When summary item is clicked, a table summarizing the results of all the pathways is shown in the lower section. Nevertheless, if a pathway is clicked, the results of this pathway are shown. These individual results are different depending on the test selected. If multilevel modeling is selected, the results consist of a table indicating the mean and the confidence interval of each subpathway in the pathway, a graphical representation of this table and a pathway context representation of the results. On the other hand, if the Wilcoxon test is selected, the results consist of a table indicating the p-value and its FDR-corrected value for each subpathway in the pathway and a representation of the pathway context results.

Chapter

5

Other applications of the method

"Science never solves a problem without creating ten more."

— George Bernard Shaw

The methodology developed in this thesis was used to the develop of another two tools: PATHiPRED[115] and PATHiVAR[143].

PATHiPRED (Section 5.1) is a web-tool for predicting discrete classes as well as continuous variables from a microarray expression dataset. This web-tool allows the user to create biomarkers associated with a certain disease and predictors of continuous variables such as drug concentrations. This approach was developed jointly with Dr. Alicia Amadoz as another application of the methodology developed in this thesis. In this case, instead of comparing of two classes, the modeling and subpathways probabilities estimation was used to create discrete and continuous variable predictors.

PATHiVAR (Section 5.2) is a web-tool for visualizing and analyzing sequencing data in a pathway context. This tool allows the user to look for the subpathways affected by a certain genomic variation taking into account the heredity assumed model and how harmful the variant is. This approach was developed in the context of Rosa Divina Hernansaiz-Ballesteros Biotechnology's research project (RP) [144] and later in her Bioinformatics master's degree research project (MRP)[145], both directed by Dr. Joaquin Dopazo and myself. The approach uses the pathway modeling and the subpathway signal transmission to asses the impact of genomic variants on different functions.

5.1 Predictors from sub-pathway probabilities

5.1.1 Abstract

The accumulation of a huge number of microarray datasets covering several diseases aver the last decade has fostered in an active quest for disease biomarkers. Initially, these biomarkers were genes selected according to their ability to discriminate among the classes compared [146, 147, 148]. However, the biomarkers found are not always reproducible across different studies [149, 150, 151], usually because complex traits are not produced by only one gene, but by a set of genes that jointly causes the disease [149, 152]. The set of genes implied in the disease can vary across different individuals, but the loss/gain of functionality they provoke has to be shared between the different individuals. This premise has motivated recent approaches that propose subnetworks as biomarkers with better predictive power than single-gene biomarkers [153]. Some of these approaches have produced effective biomarkers for different diseases such as: diabetes [154], Alzheimer [155] and different types of cancer as breast cancer [156, 157, 158], ovarian cancer [159, 160, 161], glioblastoma [159, 87] or prostate cancer [162]. These sub-network structures have also been used to predict treatment responses [163] as well as to search for drug targets [164].

The approach presented here dissects pathways into subpathways and checks these as possible biomarkers for distinguishing between different classes (discrete variables) or to predict a continuous variable. In addition, in the same way as the PATHiWAYS approach, expression values are transformed into activation probabilities in order to make measurements between genes comparable. That is, the measurement used to perform predictions in this platform is the probability that a subpathway is active.

5.1.2 Material and Methods

Two approaches can be distinguished according the type of variable being predicted: class prediction (discrete variables) and continuous variable prediction. The input to both methods is the activation probability of each subpathway considered. To calculate these probabilities from expression data they were normalized using RMA, as explained in 3.2.1. The probability of subpathway activation was then estimated following the work-flow developed in this thesis (See Figure 3.1). In this case, the subpathway probabilities are not comparable

between each other because of the high number of nodes the low probabilities[1]. To solve this problem, the subpathway probabilities were corrected using the number of subpathway nodes to obtain comparable activation probabilities.

Class prediction

First, subpathways showing no variability across the compared classes were discarded in order not to introduce noise in the model. These subpathways cannot help to distinguish between the classes being compared because they behave the same in all the classes, so deleting them helps to reduce noise in the model as well as reducing the dimensionality of the input data and therefore the computing time.

Feature selection was then performed using the Correlation-based Feature Selection (CFS) method [165]. Class prediction was carried out using subpathway activation probability matrices. The algorithms used for prediction are: K-Nearest neighbor (KNN), Random Forest (RF) [166] and Support Vector Machine (SVM) [116] and the parameters used to evaluate the classifications obtained were: accuracy, Mathews correlation coefficient (MCC), root mean square error (RMSC) and the area under the curve (AUC); all of these methods were implemented in the Babelomics platform [136].

To test our approach, datasets for two cancers were employed, both downloaded from the Gene Expression Omnibus (GEO) public repository at the National Center for Biotechnology Information (NCBI) [26]. The first one (GSE9476) was related to acute myeloid leukemia (AML) and contained data for 26 healthy donors and 26 AML patients [114]. The second one was a breast cancer dataset (GSE27562) composed of 31 normal samples and 31 malignant peripheral blood mononuclear cells (PBMC) samples[113].

Continuous variable prediction

In continuous variable prediction, subpathways which do not vary are not deleted, and so all the subpathway probabilities are used to perform a SVM regression selecting the best γ and cost parameters from the different values tested (10, and 100 cost values; and $10^(-6)$, $10(^-5)$, $10^(-4)$, $10(^-3)$ γ values), by optimizing the mean squared error of the model with a 10-fold cross-validation. If required, feature selection was performed using the CFS method [165].

[1]This point was not a problem in PATHiWAYS approach, because in that case the subpathway probabilities were compared 2 at time and so no bias caused by the number of nodes belonging to the subpathway could exist.

To test this approach, studies which give the concentration at which drug responses reach 50% (IC_{50}) absolute inhibition were used. Two large datasets were used for this purpouse, one from the Cancer Genome Project (CPG) [167] (E-MTAB-783) and the other from the Cancer Cell Line Encyclopedia (CCLE) [168] (GSE27562). These datasets provide information about the IC_{50} variable for 138 drugs and 661 cell lines which correspond to 17 cancers in the case of CPG and 24 drugs and 493 cells lines corresponding to 24 cancers in the case of CCLE.

5.1.3 Results

Use of signaling circuit activation probabilities in the context of prediction

In order to determine how useful these mechanism-based biomarkers are for prediction purposes, these prediction algorithms where tested in the context of gene expression in the AML and breast cancer datasets (Subsection 5.1.2). The data were normalized and transformed into subpathway activation probabilities as previously described (Section 3.2). Next, feature selection was carried out using CFS and the selected biomarkers were used to predict the class (case or control) to which samples in each dataset belong to. The K-Nearest neighbor (KNN), Random Forest (RF) and Support Vector Machine (SVM) algorithms were used for this analysis, and the accuracy of the prediction was evaluated using: accuracy, Mathews correlation coefficient (MCC), root mean square error (RMSC) and the area under the curve (AUC). The results are shown in table 5.1. All the methods gave different accuracy results but showed an excellent potential to predict the proposed mechanism-based biomarkers. However, SVM produces the best predictions in all of the accuracy parameters considered, therefore it was used to carry out all the subsequent predictions, using the activation status of the signaling circuits as features. In addition SVM also has the advantage that it can predict continuous parameters.

Performance of the classification method using mechanism-based biomarkers

To demonstrate that mechanism-based biomarkers are an efficient way to predict discrete classes, a scenario in which no discrete variables (classes) exist was used to test the predictor's ability to differentiate between two (inexistent) classes. Two different approaches were tested: 1)taking a random sample and

Dataset		KNN	RF	SVM
Breast cancer	Accuracy	0.86-0.89	0.90-0.92	0.99
	MCC	0.74-0.79	0.81-0.83	0.98
	RMSC	0.29-0.32	0.31	0.04
	AUC	0.97-0.98	0.98	0.99
AML	Accuracy	0.88-0.89	0.92-0.95	0.96
	MCC	0.76-0.78	0.84-0.90	0.92-0.93
	RMSC	0.27-0.30	0.31	0.10-0.11
	AUC	0.94	0.98	0.96

Table 5.1: **Results from the prediction algorithms used on the GBM dataset.** MCC is the Mathews correlation coefficient, RMSC is the root mean square error and AUC is the area under the curve

generating a dataset from it by adding different levels of noise; and 2) using a real dataset made up of different samples in the same condition. To generate the first dataset, a random sample from the Affymetrix Human Genome U133 Plus 2.0 Array dataset used to model the expression of the associated probesets (Section 3.2.1), was selected. Random noise was then added to generate N different samples, which were randomly assigned to two classes. 1000 simulations were carried out with different noise levels (SD= 0.1, 0.075, 0.05, 0.025 and 0.01 and different sample sizes (N=20, 50, 100 and 200). The real dataset was the same dataset used to perform the false positives analysis (Section 4.3), and consisted of 237 samples of Acute Myeloid Leukemia (AML)[110]. Figure 5.1 shows the distribution of accuracy values obtained in both simulated scenarios. In both cases the average value is slightly over 0.5, i.e. equivalent to a random choice between two equal possibilities. Therefore, the ratio of false positives can be considered negligible.

Prediction of IC_{50} values for cancer drugs

Because mechanism-based biomarkers are efficient at predicting discrete classes, we also explored their efficiency in predicting a continuous value. To test how oue approach behave when predicting continuous variables, the CPG dataset was used to train the predictors and CCLE dataset was used to validate them (Section 5.1.2). Both datasets were filtered to obtain data a total of 317 cell lines, 12 cancers and 7 drugs that are common them both. The CGP data were

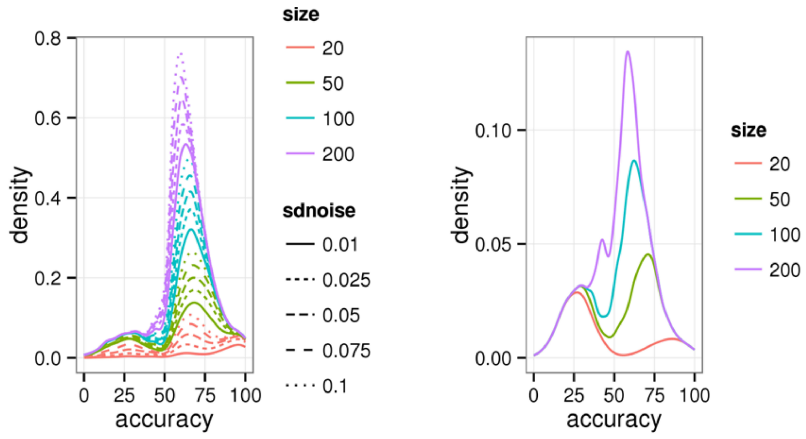


Figure 1

Figure 5.1: **Estimation of the false positives ratio.** Accuracy (cross-validation) for the two simulated classes (discrete variable) scenarios: A) classes made with randomly partitioning a pediatric AML dataset of identical AML patients and B) simulated array probes. The cross-validation values are slightly higher than 0.5, that is, indicating correctAccuracy (cross-validation) for the two scenarios of simulated classes: A) classes made with random partitions of a pediatric AML dataset of identical patients and B) simulated array probes. The cross-validation values are slightly higher than 0.5, indicating correct class assignation by pure chance.

normalized and these normalized gene expression values were transformed into signaling subpathway activation probabilities as previously described. Finally, a predictive model was obtained for the CGP data after SVM-regression, and this data was used to predict the IC50 values of the CCLE dataset. Figure 5.2 shows the agreement between the predicted a real values. There was a highly significant positive correlation (r=0.709, p=8.98x10-193) between the predicted

Cancers	Paclitaxel	AZD6244	Nilotinib	PLX4720	Sorafenib	Erlotinib	Lapatinib
Lung	2.99	3.59	2.77	3.85	2.82	5.08	3.87
Haematopoietic-lymphoid	2.98	2.68	2.42	3.33	3.05	4.88	3.89
Bone	3.96	3.01	NA	2.5	2.78	3.61	3.96
Skin	1.63	4.71	2.02	5.3	1.19	2.66	3.22
Ovary	NA	3.33	3.18	3.77	NA	NA	NA
Central nervous	2.67	4.51	2.83	3.66	2.29	2.41	2.47
Pancreas	NA	2.08	4.34	3.85	NA	NA	NA
Soft tissue	0.64	3.01	2.88	2.88	3.77	3.99	3.59
Breast	2.23	3.87	3.17	NA	3.67	5.19	5.28
Upper aerodigestive tract	0.52	2.3	2.73	NA	1.86	2.17	1.49
Kidney	1.22	NA	2.88	NA	NA	NA	NA
Thyroid	NA	2.51	2.58	1.71	NA	NA	NA

Table 5.2: **RMSE per cancer type and drug in the CCLE dataset.** NAs appears when not enough data were available.

and real IC50 values measured in the CCLE dataset which clearly proves the validity of the prediction framework we propose here. Figure 5.3 shows the predicted IC50 values and the corresponding real IC50 values available for the CCLE dataset averaged by tissue. Both the predicted and real IC50 values were compared by estimating the root mean square error (RMSE) (Table 5.2). While there are some discrepancies, the global RMSE of 3.31,which includes all cancers and drugs, demonstrates that the prediction accuracy was reasonable. Specific cancers and drugs for which the prediction is especially good are: the upper aerodigestive tract (RMSE=0.52) and soft tissue (RMSE=0.64) treated with Paclitaxel.

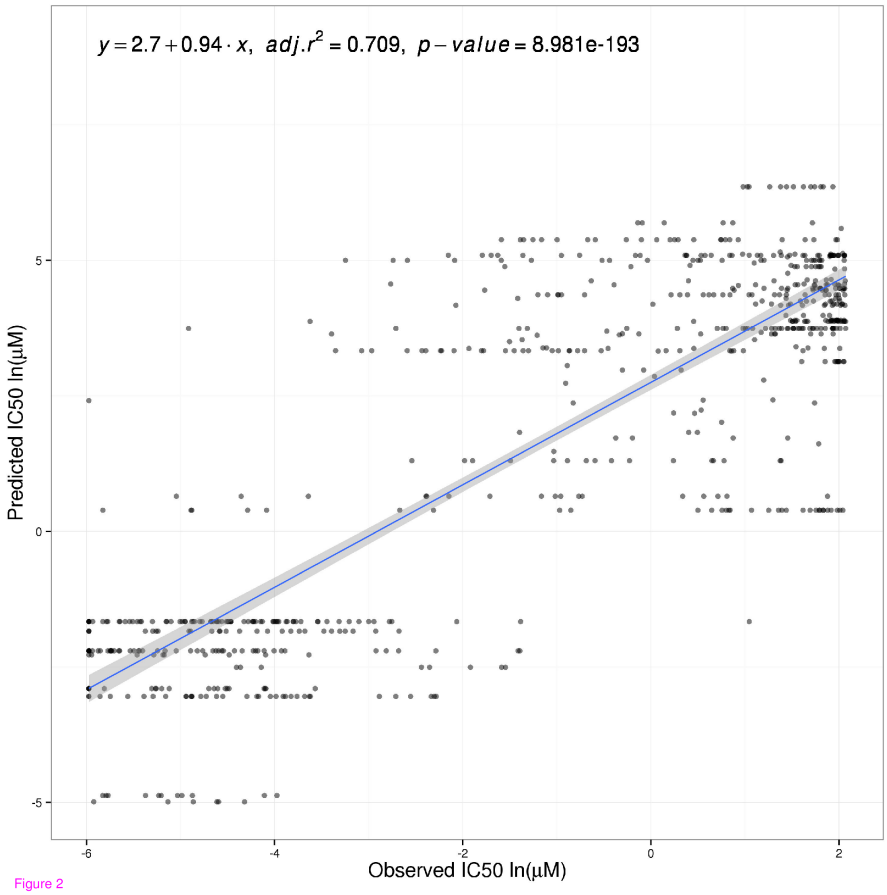


Figure 5.2: **Predicted versus actual CCLE IC50 values.**

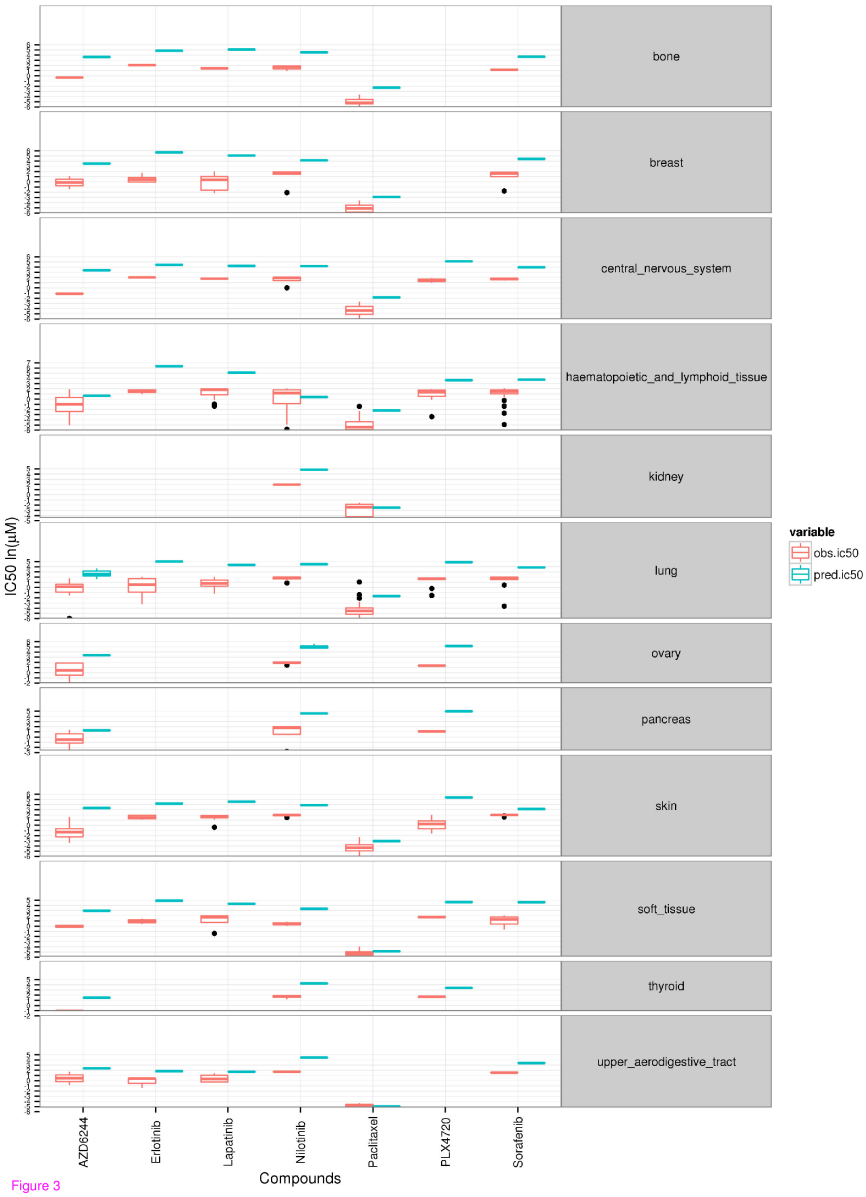


Figure 3

Figure 5.3: **CCLE predicted and observed IC50 values per cancer and drug.**

5.1.4 PATHiPRED tool

In the same way as for PATHiWAYS, the developed methodology was implemented in a web-tool called PATHiPRED[115] located at the same URL as PATHiWAYS[2]. PATHiPRED differentiates between two classes or predicts a continuous variable using the activation probabilities from stimulus-response sub-pathways calculated by PATHiWAYS methodology. PATHiPRED performs a SVM modelling with cross-validation and the results include details about the prediction model, the statistical parameters used to assess the goodness of the model, the confusion matrix of the prediction, the activation probabilities matrix, the mechanism-based biomarkers and pathways graphs with the selected sub-pathways highlighted. Moreover, the prediction model obtained can be applied to a new dataset within PATHiPRED web tool.

The program use CEL files that are transformed into subpathway activation probabilities as input to derive predictions about conditions under study. Additionally, a matrix (samples x probabilities) can be saved which can subsequently used in other programs to derive predictors with other algorithms (for example in Babelomics[136]), if desired.

5.1.5 Prediction of molecular phenotypes - The Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenge

DREAM (Dialogue for Reverse Engineering Assessments and Methods) science challenges are a non-profit and open effort to pose fundamental questions about systems biology and translational medicine. Participants propose different solutions to a common problem and all the contributions are shared to improve everyone's scientific knowledge. We participated in the DREAM Toxicogenetics Challenge to build models of cytotoxicity as mediated by exposure to environmental toxicants and drugs. In our case, DREAM provided a dataset containing estimations of cytotoxicity as measured in lymphoblastoid cell lines derived from 884 individuals following in vitro exposure to 156 chemical compounds (Figure 5.4). Given these data, two different subchallenges were given:

Subchallenge 1 Predict interindividual variability in in vitro cytotoxicity based on genomic profiles of individual cell lines. For each compound, participants will be challenged to predict the absolute values and relative

[2]Just a bottom to run PATHiPRED was added to the same web-tool

ranks of cytotoxicity across a set of unknown cell lines for which genomic data is available.

Subchallenge 2 For each compound, predict the concentration at which median cytotoxicity would occur, as well as inter-individual variation in cytotoxicity, described by the 5-95th%ile range, across the population. Each prediction will be scored based on the participant's ability to predict these two parameters within a set of compounds excluded from the training set.

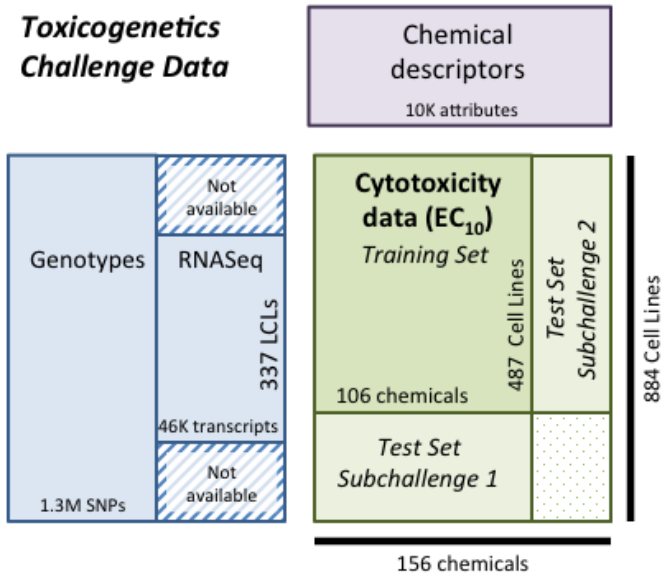


Figure 5.4: **Data provided to DREAM Toxicogenetics Challenge**

We participated in the second subchallenge. The data was used to estimate the probabilities in signal transduction for all the stimulus-response circuits of each signaling pathway starting from gene expression values. These probabilities were used in a SVM ϵ-regression with k-fold cross-validation to obtain a prediction model. Only samples for which RNA-seq data is available were used for this predictive approach. This data measures the quantity of mRNA present in the sample, in the same way that microarrays do. However, in this

case it was not possible to define the background data because there were not enough data available. Therefore, we transformed the RNA-seq data in a [0,1] interval, where genes with higher mRNA expression have a higher value in the transformation and vice versa. Thus, using this value as a probability, we estimated the probabilities of signal-transduction subpathways activation. These estimated probabilities were used as mechanism-based biomarkers in the prediction algorithm instead of the raw gene expression values.

Using the approach proposed in this thesis, the expression values of 2,585 genes included in human signaling pathways were recoded into 1,126 signal transduction probabilities, corresponding to all the signaling circuits that can be defined in the 26 pathways analyzed.

RNA-seq count data were available for 192 individuals, using these data, we fitted a regression model for each drug. First, circuits whose activation probability was constant across all individuals were discarded as potential biomarkers. Then, the remaining biomarkers were filtered according to each compound by CFS using the FSelector R library [165]. Next, SVM ϵ-regression we performed with the selected circuits using the e1071 R library [169] selecting the best γ and cost parameters from among the different values tested (10 and 100 for cost values; 10^{-6}, 10^{-5}, 10^{-4} and 10^{-3} for γ values), by optimizing the mean squared error of the model with a 10-fold cross-validation.

The results uploaded by different teams were ranked according to different metrics:

- The mean ranking PC: average of the rankings computed using PC based on the predicted versus observed median response (PCm) and the predicted versus observed interquantile distance (PCq)
- The rank PC: rank of mean ranking PC
- The mean ranking SC: average of the rankings computed using SC based on the predicted versus observed median response (SCm) and the predicted versus observed interquantile distance (SCq)
- The rank SC: rank of mean ranking SC
- The mean ranking: average of mean ranking PC and mean ranking SC
- The rank: rank of mean ranking

According to these metrics, our team was located 85/99. A limitation of our NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge prediction results

was that RNA-seq data were not available for 167 individuals in the submission dataset. Consequently, our global results were penalized in the scoring criteria because more than a half of the final samples were missing. However, the subset of 97 individuals for which RNA-seq data were available had a root mean square error (RMSE) lower than 0.3, which was obtained in the cross-validation process of the best models calculation.

The results of this challenge were published in Nature Biotechnology as a paper describing the results analysis and broadly applicable insights that arose from the DREAM Toxicogenetics Challenge [170].

5.2 Pathway analysis of structural genomic data

5.2.1 Biological introduction

A biological introduction is first required here because this approach is built upon different biological concepts.

Gene expression is the final consequence of transcription of genome's information (Section 1.2). Therefore, a variation in the DNA information due a mutation can produce a change in gene expression and consequently a potential protein production failure that can provoke a disease.

In 2007, J.Craig Venter[22][3] and Watson[172] quasi-simultaneously published a complete human genome sequence. These studies were followed by another publications that jointly increased our knowledge of gene variation[173]. Analysis of these human genomes gave rise to two main conclusions: variants are more common than expected and the association between a variant and a determined pathology is not always straightforward.

Variations in the genome are identified using DNA sequencing techniques. Techniques for genome sequencing have taken great strides forward to allow a huge quantity of samples in a relatively short period of time. These advances have motivated projects to sequence a huge number of individuals (or populations) who share a particular characteristic such as certain disease (e.g different types of cancer).

Molecular medicine studies variations in the genome that provoke diseases. These are defined as modified versions (from one to several nucleotides) when compared to the original gene. There are different types of variations in the genome (Figure 5.5).

[3] In the context of Human Genome Project[171] (HGP)

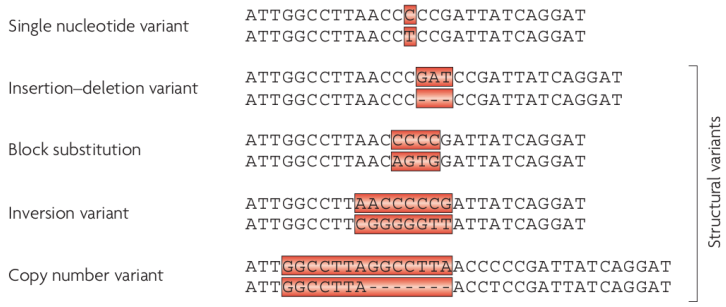


Figure 5.5: **Types of genomic variants** (Figure extracted from [173]) Single nucleotide variants are DNA sequence variations in which a single nucleotide (A, T, G or C) is altered. Insertion–deletion variants (indels) occur when one or more base pairs are present in some genomes but absent in others. Block substitutions describe cases in which a string of adjacent nucleotides varies between two genomes. An inversion variant is one in which the order of the base pairs is reversed in a defined section of a chromosome. Copy number variants occur when identical or nearly identical sequences are repeated in some chromosomes but not others.

The consequence of a variant in the production of the associated protein depends on both: the region of the gene where the variant is located (Table 5.3) and the type of aminoacid change (Table 5.4). If a variant is located in a intergenic region, the resulting protein is not affected, because only exons are transcribed, but if it is located in a exonic region, it is possible that the production of the protein will be affected. On the other hand, if the change in nucleotides does not affect the resulting aminoacid[4] (synonymous variant), the protein sequence is not affected and the protein remains functional.

In addition, there are some scores that indicate how harmful the variant is to the structure of the resulting protein:

- Sorting Intolerant From Tolerant (SIFT)[175]: Predicts whether an amino acid substitution affects protein function. Its prediction is based on the degree of conservation of amino acids residues in the sequence alignments derived from closely related sequences.

[4] See table 1.1

Annovar code	Region
UTR3	3'UTR region
UTR5	5'UTR region
exonic	coding region
splicing	splicing site
ncRNA	non-coding RNA
intronic	intronic region
upstream	region towards the 5' end of the strand
downstream	region towards the 3' end of the strand
intergenic	intergenic region

Table 5.3: **Biological consequences of the different types of variants depending on the region where the nucleotide is located**. This coding, indicated by *Annovar*[174], is used to annotated the consequence of the DNA variations.

- Polymorphism Phenotyping (PolyPhen)[176]: Predicts the possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.

Deleterious variants also depend on which heredity model is taken into account. Diploid organisms, such as humans, contain two copies of each gene, one on each strand, corresponding to the paternal and maternal alleles. Based on the occurrence of the mutation in each copy, two main heredity models can be defined: dominant and recessive. In the dominant model, a gene is considered to be non-functional if exists a deleterious variant in almost one copy[5] of the gene (heterozygous gene). Otherwise, in the recessive model, a gen is considered to be not functional if both copies of the gen have a variant (homozygous gene). In this case two different options could be considered: homozygous or compound heterozygous gene. In the first case, there must exists a variant in both copies of the gene in order to consider it as deleterious. In the case of compound heterozygous, a gene is considered as non-functional if exist almost two positions of the gene having a variant.

[5]DNA is a double stranded molecule, so it produces two copies of the gene, one for each strand

Annovar code	Aminoacid change
frameshift insertion	insertion of one or more bases that change the reading frame
frameshift deletion	deletion of one or more bases that change the reading frame
frameshift block substitution	block substitution of one or more bases that change the reading frame
stopgain	variant that provoke a stop codon creation
stoploss	variant that provoke a stop codon deletion
nonframeshift insertion	insertion of one or more bases that do not change the reading frame
nonframeshift deletion	deletion of one or more bases that do not change the reading frame
nonframeshift block substitution	block substitution of one or more bases that do not change the reading frame
nonsynonymous SNV	variant that do not cause a change in the resulting amino acid
synonymous SNV	variant that cause a change in the resulting amino acid
unknown	unknown function of the variant

Table 5.4: **Biological consequences of the different types of variants depending on the aminoacid's change**. This coding, as indicated by *Annovar*[174], is used to annotate the consequence of DNA variations.

5.2.2 Motivation

The approach presented here was developed in the context a Biotechnology's research project (RP) and later in the Bioinformatics master's degree research project (MRP), both undertaken by Rosa Divina Hernansaiz Ballesteros, and directed by Dr. Joaquin Dopazo and myself[144, 145].

The motivation for developing this approach was to study variations in healthy people from several populations in the context of signaling pathways. Their structure is used to look for differences in terms of the functionality of signaling subpathways in the different populations and to study the robustness of disrupted functional subpathways against variants in a healthy population. Therefore the objectives of this work were:

- To compare the functional consequences of variants in healthy people between different populations.

- To study the robustness of variants in terms of loss-of-functionality in signaling subpathways.
- To compare functional pathways in different tissues according to the deleterious variants that map to them.
- Implement a web-tool to visualize the predicted changes in signaling transduction probabilities across the different signaling pathways.

The tool provided from this work was recently published [143] in Nucleid Acids Research and another paper covering the first two objectives is under preparation.

5.2.3 Material and Methods

The data used in this project describe the genomes of healthy people from different populations in *1000 Genomes Project*[177] database. The *1000 Genomes Project*[177] aims to build a resource to help to understand the genetic contribution to disease and the relationship between genotype and phenotype. it includes samples from healthy donors from 14 populations that can be classified into 4 super-populations (Table 5.5)

Data was download from the *1000 Genomes Project* web-page and the functional consequences of variants were annotated by using *Annovar*[174] (See Tables 5.3 and 5.4), to obtain 14 files (one for each population) that contain information about the variants mapping all the 980 genes included in the set of selected pathways[6]. Once the variants are annotated, they have to be classified into deleterious and non-deleterious, using different filters.

The initial work (RP) only took the functional consequences of the variants and the heredity model considered into account.We compared healthy populations and analyzed the robustness of the signaling pathways in the face of aleatory genome variations for the three heredity models. However, both the dominant and recessive model classification was too restrictive, and so compound heterozygous gene modelling is considered to be best represent a common hereditary model. Variants annotated as *splicing*, *stoploss*, *stopgain* and *frameshift* were considered to be deleterious variants.

The analysis performed for the MRP not only took variant information into account, but also information about gene expression in different tissues as a filter for KEGG pathway information, thus integrating both the genomics and

[6]This set is the same as the one indicated in Section 4.1

Super-population	Population	abb
Europe	Utah residents with Northern and Western European ancestry	CEU
	Toscani in Italy	TSI
	British in England and Scotland	GBR
	Finnish in Finland	FIN
	Iberian populations in Spain	IBS
Africa	Luhya in Webuye	LWK
	Yoruba in Ibadan	YRI
	African Ancestry in Southwest US	ASW
East Asia	Han Chinese in Beijing	CHB
	Japanese in Tokyo	JPT
	Southern Han Chinese	CHS
Americas	Puerto Rican in Puerto Rico	PUR
	Mexican Ancestry in Los Angeles	MXL
	Colombian in Medellin	CLM

Table 5.5: **Populations considered in the *1000 Genomes Project.*** This table shows its classification into super-populations as well as the abbreviations for these populations

trancriptomics information. Once genes are selected as expressed in the tissue, deleterious variants are selected by their functional consequence: *splicing*, *stoploss*, *stopgain*, *frameshift* and *nonsynonymous*. In this case, *nonsynonymous* variants are classified as potentially deleterious, but are only considered as deleterious if they have an associated SIFT value less than 0.05 and a PolyPhen value greater than 0.85[7].

KEGG nodes include information for more than one gene (Section 3.1.1) acting in different tissues. To asses which genes are specifically act in each tissue data from [142], including 1,033 samples, or expression data including 66 normal tissues (Figure 5.6), were used.

These data were downloaded and normalized using the RMA[14] method (See Section 1.2.1). Genes were classified into active/inactive in the tissue using the presence-absence calls with negative probesets (PANPs)[178] approach that

[7]These values are the recommended by the authors of the respective papers

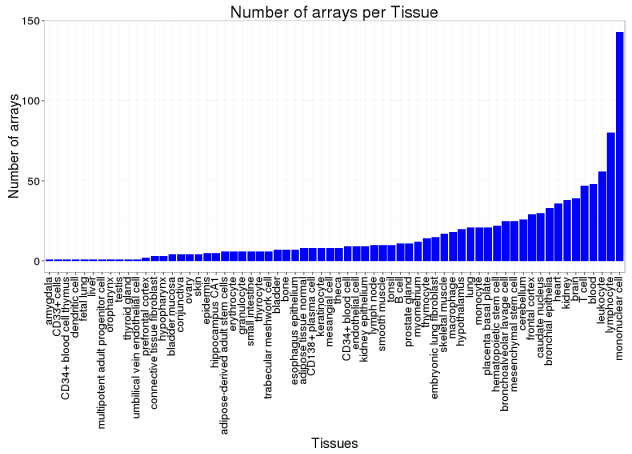


Figure 5.6: **Number of samples for each tissue** in [142]. Different tissues have a different number of samples ranging from 1 to 43.

uses PM measurements to establish the presence or absence of their activity. This filter drastically reduced the number of subpathways considered in the analysis, because only subpathways in which all the nodes were active in the tissue were took into account.

Once the variants are classified into deleterious or non-deleterious, this information is used to classify genes as functional or non-functional. A gene is defined as functional if it has a 100% probability of producing a functional protein, that is, genes are functional if they do not have any deleterious variable mapping to them. Using this coding system in the context of the methodology developed in this thesis, we could assume that if a gene is functional, it has a 100% probability of transmitting the signal along the network (so it can be considered as completely active), and if not, it can be considered as non-active with a 0% of probability of transmitting the signal along the network. Therefore,

$$p = \begin{cases} 1 & \text{Functional gene} \\ 0 & \text{Non-Functional gene} \end{cases}$$

After applying this simplification is done, the rules in section 3.4 can be applied in order to calculate the 'probability' of a node being functional and finally the 'probability' of a sub-pathway being functional. In this approach,

the 'probabilities' considered are binaries (0 or 1) so, the rules for transmitting the signal along the sub-pathways could be summarized in terms of logical operations (Figure 5.7). In this way, a node is considered as functional if any of the genes mapping to it are considered functional, a complex is considered functional if all nodes in it are functional, a path is considered functional if all nodes and complexes that compose it are functional and finally, a subpathway is considered functional if any of the paths comprise it are functional.

Figure 5.7: **Pipeline for propagating binary signals through a sub-pathway.** It is based on gene classification into deleterious or non-deleterious, functional or non-functional, and finally active or inactive in the tissue samples download from the 1000 Genomes Project using the presence-absence calls with negative probesets (PANPs) approach to make PM measurements.

These rules were applied for each individual in each population in order to dissect if each sub-pathway should be considered as functional or not, to produce a binary matrix summarizing the information about the 656 sub-pathways in the 1,167 individuals of the 14 populations considered[8]. Once this matrix is obtained this information can be used in two different approaches:

1. Comparison between healthy populations in terms of functionality of the sub-pathways

2. Evaluation of the functionality of the different sub-pathways against random variants.

We compared the populations using the Barnard test [**?**] on all of the sub-pathways and by correcting by multiple testing with the False Discovery Rate (FDR). Evaluation of the robustness of the sub-pathways against random variants was done by simulating random variants in the 980 genes under study by using the frequency of variation as in the real data multiplied by 100 times. This simulated data is put through the pipeline shown in Figure 5.7 to obtain 100 simulated sub-pathway functionality matrices. The results obtained are graphically then compared with the real ones.

[8] In the case of the MRP this analysis was done 66 times, one for each tissue.

5.2.4 Results

Comparison between healthy populations

As an example of the differences between populations in terms of functional sub-pathways, some sub-pathways in the *Melanogenesis Signaling Pathway* appear to be differentially functional between African, Asian and European populations (Figure 5.8): YRI individuals has the orange subpathway as not functional and the green one appears as non-functional in comparison between african and asiatic individuals. Genes in these sub-pathways are included in a study[179] that relates them to different pigmentation phenotypes in the different populations.

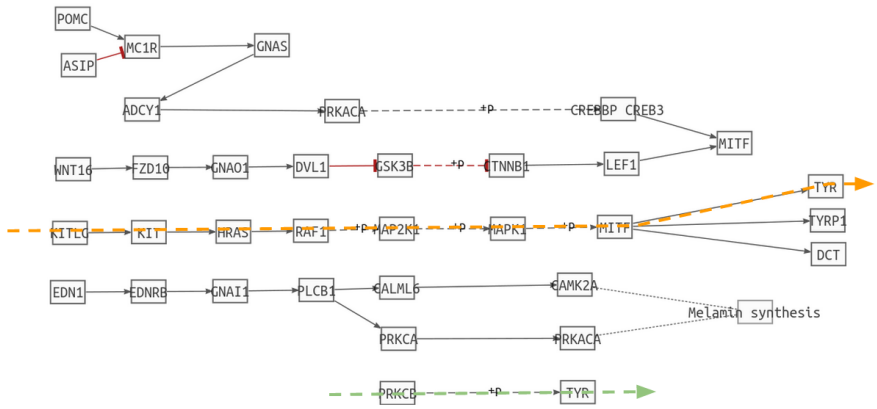
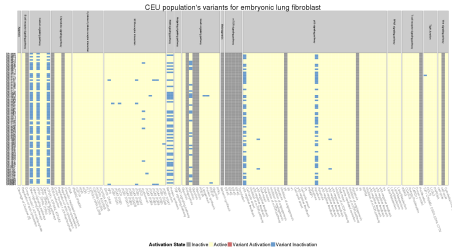


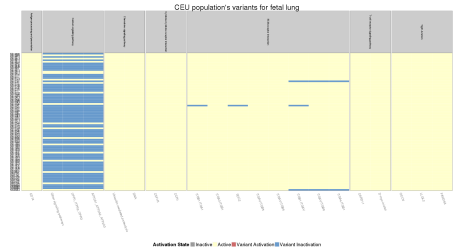
Figure 5.8: **Sub-pathways that are differentially functional between different super-populations** derived from the 1000 Genomes Project in the because of the presence of variants in the *Melanogenesis* pathway. YRI individuals has the orange subpathway as not functional and the green one appears as non-functional in comparison between african and asiatic individuals.

Comparison between different tissues

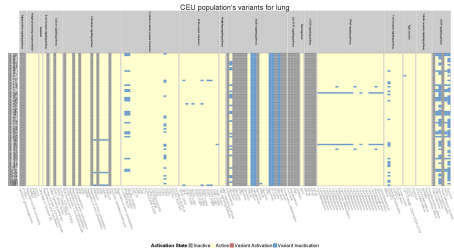
An interesting result came from analyzing of the behavior of the different subpathways in the context of tissue development. For example, we compared the expression patterns and variants in lung tissue from three different differentation stages: embryonic, fetal and adult (Figure 5.9), and showed that different subpathways were functional at different stages of development, thus highlighting how signals change throughout the development processes



(a) Embrionic Lung



(b) Fetal Lung



(c) Adult Lung

Figure 5.9: **Evolution of lung tissue in terms of functional subpathways**

Pathways' robustness

Regarding to the robustness of sub-pathways against random variants, the study was done in all populations giving similar results. In Figure 5.10 results with CEU population are shown. Results with real data are represented in red and results using simulated data are represented in black (join with their confidence intervals). We can observe that there are pathways which have all its subpathways non-functional (e.g.*Notch signaling pathway*) and other which have all its subpathways functional (e.g. *Fc Epsilon Ri signaling pathway*). From this analysis we can conclude that the functionality of the subpathways changes when comparing real and simulated data which have random variants. As an example, in the *Fc Epsilon Ri signaling pathway*, real data show all subpathways as functional, but simulated data shows a lower functionality in the same subpathways.

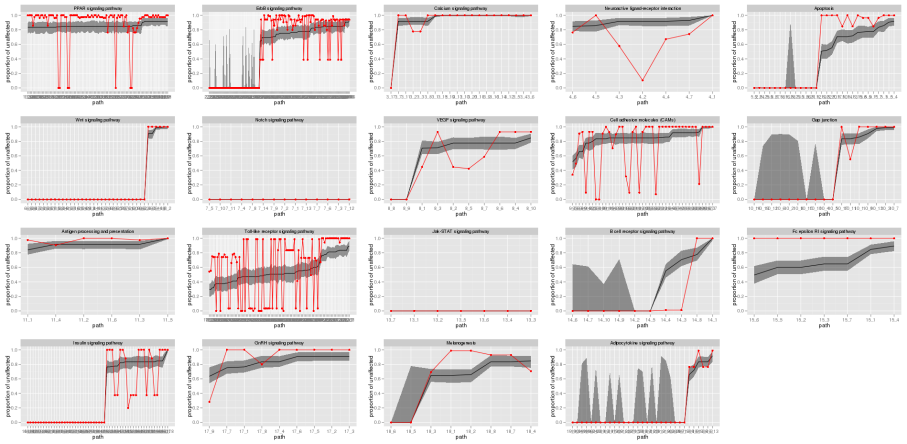


Figure 5.10: **Results of simulations** compared with real data considering recessive with compound heterozygous heredity model. Results with real data are represented in red and results using simulated data are represented in black (join with their confidence intervals)

PATHiVAR tool

The PATHiVAR web-tool[143] (Figure 5.11) estimates the functional impact that mutations have over the human signaling network. PATHiVAR analyses VCF files, extract the deleterious mutations, locates them in the signalling pathways in the selected tissue (with the appropriate expression pattern) and provides a comprehensive, graphic and interactive view of the predicted signal transduction probabilities across the different signalling pathways.

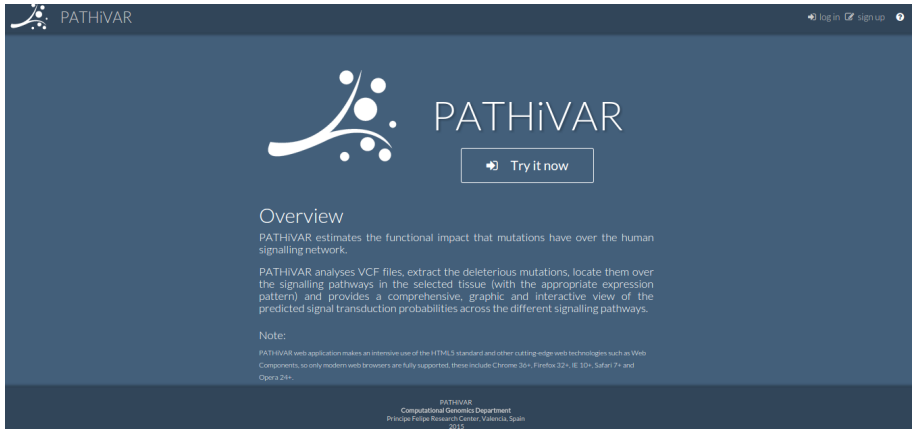


Figure 5.11: **Main page of PATHiVAR analysis tool**

The output of this tool is a table indicating all the pathways whose function was significantly affected in their functionality as well as a representation of these data in a pathway context in the same way the PATHiWAYS tool does.

CHAPTER

6 Discussion

Understanding the role of genetics in disease is now a hot topic in medical research. Some diseases are provoked by a single gene, but other complex diseases, such as cancer, are caused by the malfunction of a set of genes although the specific genes involved can change depending on the individual affected. This can be explained because different genes share functionality and/or are implicated in the same pathway. The approach presented on this thesis directly relates to expression data, i.e. measuring genotype, by the gain/loss of cell functionalities in a pathway context. This allows researchers to relate the changes in gene expression due to a specific diseases, with changes in the pathway functionality.

The study of gene expression data from control-case experiments in a pathway context is broadly studied by Bioinformatics. There are several methods that estimate the impact of a certain condition on a pathway context[1]. These methods can be divided into four types, or generations, where new generation has been developed in order to overcome the limitations of the previous ones.

The most recent set of methods are termed subpathway-based approaches and the method developed here is included in this generation. SB-methods solve most of limitations of the previous methods; they use expression data as well as pathway structure to estimate the impact of a disease, taking into account the relationships established between the different elements of the pathway. They do not describe the impact of the whole pathway, but rather, they define pathway substructures called subpathways. These structures allow the researcher to focus on the specific part of the network affected by the disease. However, these methods do still present some limitations, and the approach presented in this thesis is attempt to address these issues.

The first limitation is a result of the way pathways are modeled. The pathways in the KEGG contain information about some specific characteris-

[1] For a detailed explanation of pathway analysis methods see section 1.5

tics that must be taken into account in the analysis of this data, as described in on section 3.1. There are two concepts that are usually dismissed in almost all pathway analysis methods: groups of genes that encode scaffold proteins, and the different nature of the relationships established between gene network nodes. Groups are considered as separate nodes in the current pathway analysis methods. However, they represent a set of nodes that have to act together to propagate the signal, i.e. they are considered as independent when in fact they have to act together for the signal to pass. Regarding the nature of the different relationships, they can be classified as activations and inhibitions, and the consequence of each of these is different in terms of signal propagation. Activators enhance the proliferation of the signal along the network while inhibitors stop it. Therefore, not considering the different effects in that these relationships have on either enhancement or diminishment of signal propagation (as in most current pathway analysis methods) creates error.

Another limitation of the previous methods is the lack of interpretability of the results obtained. The jump from studying pathways to studying subpathways came from the attempts to accounts for the fact that pathways are structures which do not have well-defined limits, i.e. each database defines pathways from different points of view, so the boundaries between them are arbitrary. In the KEGG pathway database, with several different functionalities are associated with the same pathway, for example, the *Apoptosis* pathway includes subpathways that point both to apoptosis and to the cell cycle, which are clearly opposing functions. Therefore, defining subpathways as cliques or sub-graphs, which include several terminal nodes, produces a results which includes several and opposite functions included in the same subpathway, thus making a clear interpretation of the data obtained from them impossible.

Finally, it is very important to give the methodology developed in this thesis to the research community in an user-friendly form. Most of the previous methods provide their approach in R language, which drastically reduces the number of researches that are able to use it, because it is too technical for users who are not experts in this specific field.

All of these limitations have been overcome by our method: we model KEGG pathways according their underlying biology and; different relationships are taken into account when calculating the signal propagation along the network. We also created a user-friendly web-tool PATHIWAYS, which is provided in the following website `http://pathiways.babelomics.org/`. Another two tools, PATHiPRED and PATHiVAR, which model pathways information and how the signal is transmitted along the network in the same sense that the approach proposed here, show the utility of this modeling.

Nevertheless, the proposed approach still presents some limitations. The most importnat of which is that only a few signaling pathways can be modeled. This is, mainly because the existence of loops, and because metabolic pathways cannot be modeled in this way. Therefore in future work it is important to take this loops into account when finding ways to model how signal is transmitted along the network. This not an easy problem, because, by definition, this type of data analysis considers data from a fixed time point, but loops represent dynamic behaviors. Another limitation of this method is that the estimated subpathway probabilities lack biological meaning, they only make sense for the purposes of comparison. The different lengths of the paths and the methods of signal transmission along the network make it difficult to interpret the raw estimated probability.

In conclusion, the method developed and described in this thesis improves upon the existing methods for subpathway analysis in many ways, but still has some limitations. Thus, further work will be required to overcome these limitations. Future work should also aim to include more microarray platforms and pathway databases, perhaps even to give the user the opportunity to upload their own network annotations. Finally, future work building upon the method presented here chould aim to improve the test used to compare the data so that is more compatible with this type of data and includes the structure of the data used in its calculation algorithm.

Chapter

7 Conclusion

- **Probe set behavior for six Affymetrix microarray platforms were modeled.** Concretely, three platforms for human, two for mouse and one for rat were used. This database for probe set behaviors is the core to calculate probe set activation probabilities.
- **Pathway models representing the underlying biology were obtained from KEGG database.** This task includes the definition of proper biologically oriented definitions of the different types of both, nodes and edges. Also, sub-pathway definition oriented to functionalities provided by KEGG is a clear advantage of the proposed model.
- **A new method for analyzing sub-pathway differential activation behavior between two conditions was developed.** This approach, called PATHiWAYS, overcomes most of the limitations of previous methodologies. PATHiWAYS can be used to estimate which subpathways in a pathway network are differentially more active/inactive between two conditions by using expression data. This new methodology probed its utility with the analysis of real data.
- **A web-tool was implemented to analyze user data and provide a easy graphical interpretation in a pathway context.** This web-tool allows to upload both raw or normalized data and provides numerical and graphical results of the comparison between to experimental conditions.

Bibliography

[1] J. D. Watson, F. H. Crick, *et al.*, "Molecular structure of nucleic acids," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[2] F. Crick *et al.*, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *et al.*, *Control of gene expression*. Garland Science, 2002.

[4] J. Sambrook, "Adenovirus amazes at cold spring harbor.," *Nature*, vol. 268, no. 5616, p. 101, 1977.

[5] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, "Correlation between protein and mrna abundance in yeast," *Molecular and cellular biology*, vol. 19, no. 3, pp. 1720–1730, 1999.

[6] D. Greenbaum, C. Colangelo, K. Williams, M. Gerstein, *et al.*, "Comparing protein abundance and mrna expression levels on a genomic scale," *Genome Biol*, vol. 4, no. 9, p. 117, 2003.

[7] S. Efroni, C. F. Schaefer, and K. H. Buetow, "Identification of key processes underlying cancer phenotypes using biologic pathway analysis," *PLoS One*, vol. 2, no. 5, p. e425, 2007.

[8] J. C. Alwine, D. J. Kemp, and G. R. Stark, "Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes," *Proceedings of the National Academy of Sciences*, vol. 74, no. 12, pp. 5350–5354, 1977.

[9] K. B. Mullis and H. A. Erlich, "Primer-directed enzymatic amplification of dna with a thermostable dna polymerase," *Science*, vol. 239, no. 4839, pp. 487–491, 1988.

[10] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome research*, vol. 18, no. 9, pp. 1509–1517, 2008.

[11] A. Oshlack, M. D. Robinson, M. D. Young, *et al.*, "From rna-seq reads to differential expression results," *Genome Biol*, vol. 11, no. 12, p. 220, 2010.

[12] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Norton, *et al.*, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature biotechnology*, vol. 14, no. 13, pp. 1675–1680, 1996.

[13] J. Quackenbush, "Microarray data normalization and transformation," *Nature genetics*, vol. 32, pp. 496–501, 2002.

[14] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of affymetrix genechip probe level data," *Nucleic acids research*, vol. 31, no. 4, pp. e15–e15, 2003.

[15] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proceedings of the National Academy of Sciences*, vol. 98, no. 1, pp. 31–36, 2001.

[16] I. V. Yang, E. Chen, J. P. Hasseman, W. Liang, B. C. Frank, S. Wang, V. Sharov, A. I. Saeed, J. White, J. Li, *et al.*, "Within the fold: assessing differential expression measures and reproducibility in microarray assays," *Genome Biol*, vol. 3, no. 11, pp. 1–0062, 2002.

[17] R. Mei, X. Di, T. Ryder, E. Hubbell, S. Dee, T. Webster, C. Harrington, M. h Ho, J. Baid, S. Smeekens, *et al.*, "Analysis of high density expression microarrays with signed-rank call algorithms," *Bioinformatics*, vol. 18, no. 12, pp. 1593–1599, 2002.

[18] S. Efroni, L. Carmel, C. G. Schaefer, and K. H. Buetow, "Superposition of transcriptional behaviors determines gene state," *PloS one*, vol. 3, no. 8, p. e2901, 2008.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer New York:, 2001.

[20] N. M. Luscombe, D. Greenbaum, and M. Gerstein, "What is bioinformatics? a proposed definition and overview of the field," 2001.

[21] F. Collins, E. Lander, J. Rogers, R. Waterston, and I. Conso, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, 2004.

[22] S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, *et al.*, "The diploid genome sequence of an individual human," *PLoS biology*, vol. 5, no. 10, p. e254, 2007.

[23] R. R. Copley, "The animal in the genome: comparative genomics and evolution," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1496, pp. 1453–1461, 2008.

[24] T. Attwood, A. Gisel, N. Eriksson, and E. Bongcam-Rudloff, "Concepts, historical milestones and the central place of bioinformatics in modern biology: a european perspective," *Bioinformatics—Trends and Methodologies*, 2011.

[25] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 2002.

[26] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, *et al.*, "Ncbi geo: archive for functional genomics data sets—update," *Nucleic acids research*, vol. 41, no. D1, pp. D991–D995, 2013.

[27] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in kegg," *Nucleic acids research*, vol. 42, no. D, pp. D199–D205, 2014.

[28] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[29] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," *PLoS computational biology*, vol. 8, no. 2, p. e1002375, 2012.

[30] J. J. Goeman and P. Bühlmann, "Analyzing gene expression data in terms of gene sets: methodological issues," *Bioinformatics*, vol. 23, no. 8, pp. 980–987, 2007.

[31] C. Chen, M. T. Weirauch, C. C. Powell, A. C. Zambon, and J. M. Stuart, "A search engine to identify pathway genes from expression data on multiple organisms," *BMC Systems Biology*, vol. 1, no. 1, p. 20, 2007.

[32] C. Li, X. Li, Y. Miao, Q. Wang, W. Jiang, C. Xu, J. Li, J. Han, F. Zhang, B. Gong, *et al.*, "Subpathwayminer: a software package for flexible identification of pathways," *Nucleic acids research*, vol. 37, no. 19, pp. e131–e131, 2009.

[33] L. Jacob, P. Neuvial, and S. Dudoit, "More power via graph-structured tests for differential expression of gene networks," *The Annals of Applied Statistics*, vol. 6, no. 2, pp. 561–600, 2012.

[34] P. Martini, G. Sales, M. S. Massa, M. Chiogna, and C. Romualdi, "Along signal paths: an empirical gene set approach exploiting pathway topology," *Nucleic acids research*, vol. 41, no. 1, pp. e19–e19, 2013.

[35] G. Sales, E. Calura, P. Martini, and C. Romualdi, "Graphite web: web tool for gene set analysis exploiting pathway topology," *Nucleic acids research*, 2013.

[36] X. Chen, J. Xu, B. Huang, J. Li, X. Wu, L. Ma, X. Jia, X. Bian, F. Tan, L. Liu, *et al.*, "A sub-pathway-based approach for identifying drug response principal network," *Bioinformatics*, vol. 27, no. 5, pp. 649–654, 2011.

[37] S. Nam and T. Park, "Pathway-based evaluation in early onset colorectal cancer suggests focal adhesion and immunosuppression along with epithelial-mesenchymal transition," *PloS one*, vol. 7, no. 4, p. e31685, 2012.

[38] T. Judeh, C. Johnson, A. Kumar, and D. Zhu, "Teak: Topology enrichment analysis framework for detecting activated biological subpathways," *Nucleic acids research*, vol. 41, no. 3, pp. 1425–1437, 2013.

[39] W. A. Haynes, R. Higdon, L. Stanberry, D. Collins, and E. Kolker, "Differential expression analysis for pathways," *PLoS computational biology*, vol. 9, no. 3, p. e1002967, 2013.

[40] P. Khatri, S. Draghici, G. C. Ostermeier, and S. A. Krawetz, "Profiling gene expression using onto-express," *Genomics*, vol. 79, no. 2, pp. 266–270, 2002.

[41] S. Draghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz, "Global functional profiling of gene expression," *Genomics*, vol. 81, no. 2, pp. 98–104, 2003.

[42] S. W. Doniger, N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor, B. R. Conklin, *et al.*, "Mappfinder: using gene ontology and genmapp to create a global gene-expression profile from microarray data," *Genome biol*, vol. 4, no. 1, p. R7, 2003.

[43] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, *et al.*, "Gominer: a resource for biological interpretation of genomic and proteomic data," *Genome Biol*, vol. 4, no. 4, p. R28, 2003.

[44] F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo, "Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes," *Bioinformatics*, vol. 20, no. 4, pp. 578–580, 2004.

[45] T. Beißbarth and T. P. Speed, "Gostat: find statistically overrepresented gene ontologies within a group of genes," *Bioinformatics*, vol. 20, no. 9, pp. 1464–1465, 2004.

[46] G. F. Berriz, O. D. King, B. Bryant, C. Sander, and F. P. Roth, "Characterizing gene sets with funcassociate," *Bioinformatics*, vol. 19, no. 18, pp. 2502–2504, 2003.

[47] D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry, and B. Jacq, "Gotoolbox: functional analysis of gene datasets based on gene ontology," *Genome biology*, vol. 5, no. 12, p. R101, 2004.

[48] C. I. Castillo-Davis and D. L. Hartl, "Genemerge—post-genomic analysis, data mining, and hypothesis testing," *Bioinformatics*, vol. 19, no. 7, pp. 891–892, 2003.

[49] Q. Zheng and X.-J. Wang, "Goeast: a web-based software toolkit for gene ontology enrichment analysis," *Nucleic acids research*, vol. 36, no. suppl 2, pp. W358–W363, 2008.

[50] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski, and J. Galon, "Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks," *Bioinformatics*, vol. 25, no. 8, pp. 1091–1093, 2009.

[51] M. D. Robinson, J. Grigull, N. Mohammad, and T. R. Hughes, "Funspec: a web-based cluster interpreter for yeast," *BMC bioinformatics*, vol. 3, no. 1, p. 35, 2002.

[52] L. A. Martínez-Cruz, A. Rubio, M. L. Martínez-Chantar, A. Labarga, I. Barrio, A. Podhorski, V. Segura, J. L. S. Campo, M. A. Avila, and J. M. Mato, "Garban: genomic analysis and rapid biological annotation of cdna microarray and proteomic data," *Bioinformatics*, vol. 19, no. 16, pp. 2158–2160, 2003.

[53] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock, "Go:: Termfinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, 2004.

[54] B. Zhang, S. Kirov, and J. Snoddy, "Webgestalt: an integrated system for exploring gene sets in various biological contexts," *Nucleic acids research*, vol. 33, no. suppl 2, pp. W741–W748, 2005.

[55] D. Duncan, N. Prodduturi, and B. Zhang, "Webgestalt2: an updated and expanded version of the web-based gene set analysis toolkit," *BMC bioinformatics*, vol. 11, no. Suppl 4, p. P10, 2010.

[56] Z. Du, X. Zhou, Y. Ling, Z. Zhang, and Z. Su, "agrigo: a go analysis toolkit for the agricultural community," *Nucleic acids research*, vol. 38, no. suppl 2, pp. W64–W70, 2010.

[57] H. Sun, H. Fang, T. Chen, R. Perkins, and W. Tong, "Goffa: Gene ontology for functional analysis–a fda gene ontology tool for analysis of genomic and proteomic data," *BMC bioinformatics*, vol. 7, no. Suppl 2, p. S23, 2006.

[58] J. Ye, L. Fang, H. Zheng, Y. Zhang, J. Chen, Z. Zhang, J. Wang, S. Li, R. Li, L. Bolund, *et al.*, "Wego: a web tool for plotting go annotations," *Nucleic acids research*, vol. 34, no. suppl 2, pp. W293–W297, 2006.

[59] R. Vêncio, T. Koide, S. Gomes, and C. de B Pereira, "Baygo: Bayesian analysis of ontology term enrichment in microarray data," *BMC bioinformatics*, vol. 7, no. 1, p. 86, 2006.

[60] S. Zhong, C. Li, and W. H. Wong, "Chipinfo: Software for extracting gene annotation and gene ontology information for microarray analysis," *Nucleic acids research*, vol. 31, no. 13, pp. 3483–3486, 2003.

[61] D. Pan, N. Sun, K.-H. Cheung, Z. Guan, L. Ma, M. Holford, X. Deng, and H. Zhao, "Pathmapa: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for arabidopsis," *BMC bioinformatics*, vol. 4, no. 1, p. 56, 2003.

[62] H.-J. Chung, M. Kim, C. H. Park, J. Kim, and J. H. Kim, "Arrayxpath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using scalable vector graphics," *Nucleic acids research*, vol. 32, no. suppl 2, pp. W460–W464, 2004.

[63] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[64] M. Morgan, S. Falcon, and R. Gentleman, "Gseabase: Gene set enrichment data structures and methods," *R package version 1.2*, vol. 2, 2008.

[65] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, "Discovering statistically significant pathways in expression profiling studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13544–13549, 2005.

[66] Z. Jiang and R. Gentleman, "Extensions to gene set enrichment," *Bioinformatics*, vol. 23, no. 3, pp. 306–313, 2007.

[67] W. T. Barry, A. B. Nobel, and F. A. Wright, "Significance analysis of functional categories in gene expression studies: a structured permutation approach," *Bioinformatics*, vol. 21, no. 9, pp. 1943–1949, 2005.

[68] J. J. Goeman, S. A. Van De Geer, F. De Kort, and H. C. Van Houwelingen, "A global test for groups of genes: testing association with a clinical outcome," *Bioinformatics*, vol. 20, no. 1, pp. 93–99, 2004.

[69] S. W. Kong, W. T. Pu, and P. J. Park, "A multivariate approach for integrating genome-wide expression data and biological knowledge," *Bioinformatics*, vol. 22, no. 19, pp. 2373–2380, 2006.

[70] I. Dinu, J. D. Potter, T. Mueller, Q. Liu, A. J. Adewale, G. S. Jhangri, G. Einecke, K. S. Famulski, P. Halloran, and Y. Yasui, "Improving gene set analysis of microarray data by sam-gs," *BMC bioinformatics*, vol. 8, no. 1, p. 242, 2007.

[71] T. Breslin, P. Edén, and M. Krogh, "Comparing functional annotation analyses with catmap," *BMC bioinformatics*, vol. 5, no. 1, p. 193, 2004.

[72] A. Boorsma, B. C. Foat, D. Vis, F. Klis, and H. J. Bussemaker, "T-profiler: scoring the activity of predefined groups of genes using gene expression data," *Nucleic acids research*, vol. 33, no. suppl 2, pp. W592–W595, 2005.

[73] C. Henegar, R. Cancello, S. Rome, H. Vidal, K. Clément, and J.-D. Zucker, "Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes," *Journal of Bioinformatics and Computational Biology*, vol. 4, no. 04, pp. 833–852, 2006.

[74] C. Backes, A. Keller, J. Kuentzer, B. Kneissl, N. Comtesse, Y. A. Elnakady, R. Müller, E. Meese, and H.-P. Lenhof, "Genetrail—advanced gene set enrichment analysis," *Nucleic acids research*, vol. 35, no. suppl 2, pp. W186–W192, 2007.

[75] S.-B. Kim, S. Yang, S.-K. Kim, S. C. Kim, H. G. Woo, D. J. Volsky, S.-Y. Kim, and I.-S. Chu, "Gazer: gene set analyzer," *Bioinformatics*, vol. 23, no. 13, pp. 1697–1699, 2007.

[76] F. Al-Shahrour, L. Arbiza, H. Dopazo, J. Huerta-Cepas, P. Mínguez, D. Montaner, and J. Dopazo, "From genes to functional classes in the study of biological systems," *BMC bioinformatics*, vol. 8, no. 1, p. 114, 2007.

[77] R. Pandey, R. K. Guru, and D. W. Mount, "Pathway miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data," *Bioinformatics*, vol. 20, no. 13, pp. 2156–2158, 2004.

[78] B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, R. A. Lempicki, *et al.*, "David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists," *Nucleic acids research*, vol. 35, no. suppl 2, pp. W169–W175, 2007.

[79] P. Grosu, J. P. Townsend, D. L. Hartl, and D. Cavalieri, "Pathway processor: a tool for integrating whole-genome expression results into metabolic networks," *Genome research*, vol. 12, no. 7, pp. 1121–1126, 2002.

[80] J. Rahnenfuhrer, F. S. Domingues, J. Maydt, and T. Lengauer, "Calculating the statistical significance of changes in pathway activity from gene expression data," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, p. 1055, 2004.

[81] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, and R. Romero, "A systems biology approach for pathway level analysis," *Genome research*, vol. 17, no. 10, pp. 1537–1545, 2007.

[82] P. Khatri, S. Draghici, A. L. Tarca, S. S. Hassan, and R. Romero, "A system biology approach for the steady-state analysis of gene signaling networks," in *Progress in Pattern Recognition, Image Analysis and Applications*, pp. 32–41, Springer, 2007.

[83] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J. sun Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2009.

[84] A. L. Tarca, P. Kathri, and S. Draghici, *SPIA: Signaling Pathway Impact Analysis (SPIA) using combined evidence of pathway over-representation and unusual signaling perturbations*, 2011. R package version 2.8.0.

[85] A. Shojaie and G. Michailidis, "Analysis of gene sets based on the underlying regulatory network," *Journal of Computational Biology*, vol. 16, no. 3, pp. 407–426, 2009.

[86] S. Isci, C. Ozturk, J. Jones, and H. H. Otu, "Pathway analysis of high-throughput biological data within a bayesian network framework," *Bioinformatics*, vol. 27, no. 12, pp. 1667–1674, 2011.

[87] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, "Inference of patient-specific pathway activities

from multi-dimensional cancer genomics data using paradigm," *Bioinformatics*, vol. 26, no. 12, pp. i237–i245, 2010.

[88] P. Sebastian-Leon, E. Vidal, P. Minguez, A. Conesa, S. Tarazona, A. Amadoz, C. Armero, F. Salavert, A. Vidal-Puig, D. Montaner, and J. Dopazo, "Understanding disease mechanisms with models of signaling pathway activities," *BMC systems biology*, vol. 8, no. 1, p. 121, 2014.

[89] P. Sebastián-León, J. Carbonell, F. Salavert, R. Sanchez, I. Medina, and J. Dopazo, "Inferring the functional effect of gene expression changes in signaling pathways," *Nucleic acids research*, p. gkt451, 2013.

[90] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?," *Nature biotechnology*, vol. 28, no. 3, pp. 245–248, 2010.

[91] U. Consortium *et al.*, "Reorganizing the protein space at the universal protein resource (uniprot)," *Nucleic Acids Res*, vol. 40, pp. D71–D75, 2012.

[92] S. E. Alvarez, K. B. Harikumar, N. C. Hait, J. Allegood, G. M. Strub, E. Y. Kim, M. Maceyka, H. Jiang, C. Luo, T. Kordula, *et al.*, "Sphingosine-1-phosphate is a missing cofactor for the e3 ubiquitin ligase traf2," *Nature*, vol. 465, no. 7301, pp. 1084–1088, 2010.

[93] A. S. Don and H. Rosen, "A lipid binding domain in sphingosine kinase 2," *Biochemical and biophysical research communications*, vol. 380, no. 1, pp. 87–92, 2009.

[94] Y. Inagaki, P.-Y. Li, A. Wada, S. Mitsutake, and Y. Igarashi, "Identification of functional nuclear export sequences in human sphingosine kinase 1," *Biochemical and biophysical research communications*, vol. 311, no. 1, pp. 168–173, 2003.

[95] B. Schröder, C. Wrocklage, C. Pan, R. Jäger, B. Kösters, H. Schäfer, H.-P. Elsässer, M. Mann, and A. Hasilik, "Integral and associated lysosomal membrane proteins," *Traffic*, vol. 8, no. 12, pp. 1676–1686, 2007.

[96] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, "Genecards: a novel functional genomics compendium with automated data mining and query reformulation support.," *Bioinformatics*, vol. 14, no. 8, pp. 656–664, 1998.

[97] J. L. Gross and J. Yellen, *Handbook of graph theory*. CRC, 2003.

[98] J. Zhu and S. Davis, *GEOmetadb: A compilation of metadata from NCBI GEO*, 2011. R package version 1.16.0.

[99] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "affy—analysis of affymetrix genechip data at the probe level," *Bioinformatics*, vol. 20, no. 3, pp. 307–315, 2004.

[100] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, eds.), pp. 397–420, New York: Springer, 2005.

[101] P. Macdonald and with contributions from Juan Du, *mixdist: Finite Mixture Distribution Models*, 2011. R package version 0.5-4.

[102] P. Macdonald and P. Green, "User's Guide to Program MIX: an Interactive Program for Fitting Mixtures of Distributions, Release 2.3," *Ichthus Data Systems, Ont., Canada*, 1988.

[103] C. Rao, *Linear Statistical Inference and Its Applications, paperback*. 2002.

[104] W. Szpankowski, "Inclusion-exclusion principle," *Average Case Analysis of Algorithms on Sequences*, pp. 49–72, 2001.

[105] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908.

[106] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[107] A. Gelman, J. Hill, and M. Yajima, "Why we (usually) don't have to worry about multiple comparisons," *Journal of Research on Educational Effectiveness*, vol. 5, no. 2, pp. 189–211, 2012.

[108] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, no. 6, pp. 80–83, 1945.

[109] G. Csardi and T. Nepusz, "The igraph software package for complex network research. interjournal 2006," *Complex Systems*, vol. 1695, pp. 1–9.

[110] B. V. Balgobind, M. M. Van den Heuvel-Eibrink, R. X. De Menezes, D. Reinhardt, I. H. Hollink, S. T. Arentsen-Peters, E. R. van Wering, G. J. Kaspers, J. Cloos, E. S. de Bont, *et al.*, "Evaluation of gene expression signatures predictive of cytogenetic and molecular subtypes of pediatric acute myeloid leukemia," *haematologica*, vol. 96, no. 2, pp. 221–230, 2011.

[111] J. D. Sandahl, E. A. Coenen, E. Forestier, J. Harbott, B. Johansson, G. Kerndrup, S. Adachi, A. Auvrignon, H. B. Beverloo, J.-M. Cayuela, *et al.*, "t (6; 9)(p22; q34)/dek-nup214-rearranged pediatric myeloid leukemia: an international study of 62 patients," *haematologica*, vol. 99, no. 5, pp. 865–872, 2014.

[112] R. Thomas, J. M. Gohlke, G. F. Stopper, F. M. Parham, and C. J. Portier, "Choosing the right path: enhancement of biologically relevant sets of genes or proteins using pathway structure," *Genome Biol*, vol. 10, no. 4, p. R44, 2009.

[113] H. G. LaBreche, J. R. Nevins, and E. Huang, "Integrating factor analysis and a transgenic mouse model to reveal a peripheral blood predictor of breast tumors," *BMC medical genomics*, vol. 4, no. 1, p. 61, 2011.

[114] D. L. Stirewalt, S. Meshinchi, K. J. Kopecky, W. Fan, E. L. Pogosova-Agadjanyan, J. H. Engel, M. R. Cronk, K. S. Dorcy, A. R. McQuary, D. Hockenbery, *et al.*, "Identification of genes with abnormal expression changes in acute myeloid leukemia," *Genes, Chromosomes and Cancer*, vol. 47, no. 1, pp. 8–20, 2008.

[115] A. Amadoz, P. Sebastian-Leon, E. Vidal, F. Salavert, and J. Dopazo, "Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity," *In preparation.*

[116] V. N. Vapnik and V. Vapnik, *Statistical learning theory*, vol. 1. Wiley New York, 1998.

[117] M. Consortium *et al.*, "The microarray quality control (maqc)-ii study of common practices for the development and validation of microarray-based predictive models," *Nature biotechnology*, vol. 28, no. 8, pp. 827–838, 2010.

[118] Y. Hong, K. S. Ho, K. W. Eu, and P. Y. Cheah, "A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis," *Clinical Cancer Research*, vol. 13, no. 4, pp. 1107–1114, 2007.

[119] G. Calviello, F. Di Nicuolo, S. Gragnoli, E. Piccioni, S. Serini, N. Maggiano, G. Tringali, P. Navarra, F. O. Ranelletti, and P. Palozza, "n-3 pufas reduce vegf expression in human colon cancer cells modulating the cox-2/pge2 induced erk-1 and-2 and hif-1alpha induction pathway," *Carcinogenesis*, vol. 25, no. 12, pp. 2303–2310, 2004.

[120] O. Trifan and T. Hla, "Cyclooxygenase-2 modulates cellular growth and promotes tumorigenesis," *Journal of cellular and molecular medicine*, vol. 7, no. 3, pp. 207–222, 2003.

[121] H. F. Dvorak, "Vascular permeability factor/vascular endothelial growth factor: a critical cytokine in tumor angiogenesis and a potential target for diagnosis and therapy," *Journal of clinical oncology*, vol. 20, no. 21, pp. 4368–4380, 2002.

[122] H. Xiong, Z.-G. Zhang, X.-Q. Tian, D.-F. Sun, Q.-C. Liang, Y.-J. Zhang, R. Lu, Y.-X. Chen, and J.-Y. Fang, "Inhibition of jak1, 2/stat3 signaling induces apoptosis, cell cycle arrest, and reduces tumor cell invasion in colorectal cancer cells," *Neoplasia (New York, NY)*, vol. 10, no. 3, p. 287, 2008.

[123] J.-P. Spano, G. Milano, C. Rixe, and R. Fagard, "Jak/stat signalling pathway in colorectal cancer: a new biological target with therapeutic implications," *European Journal of Cancer*, vol. 42, no. 16, pp. 2668–2670, 2006.

[124] M. Parri and P. Chiarugi, "Rac and rho gtpases in cancer cell motility control," *Cell Commun Signal*, vol. 8, no. 23, 2010.

[125] X. Prieur, C. Y. Mok, V. R. Velagapudi, V. Núñez, L. Fuentes, D. Montaner, K. Ishikawa, A. Camacho, N. Barbarroja, S. O'Rahilly, *et al.*, "Differential lipid partitioning between adipocytes and tissue macrophages modulates macrophage lipotoxicity and m2/m1 polarization in obese mice," *Diabetes*, vol. 60, no. 3, pp. 797–809, 2011.

[126] M. Alligier, E. Meugnier, C. Debard, S. Lambert-Porcheron, E. Chanseaume, M. Sothier, E. Loizon, A. A. Hssain, J. Brozek, J.-Y.

Scoazec, *et al.*, "Subcutaneous adipose tissue remodeling during the initial phase of weight gain induced by overfeeding in humans," *The Journal of Clinical Endocrinology & Metabolism*, vol. 97, no. 2, pp. E183–E192, 2011.

[127] J. M. Olefsky *et al.*, "Wnt fans the flames in obesity," *Science*, vol. 329, no. 5990, pp. 397–398, 2010.

[128] D. Hu, A. Fukuhara, Y. Miyata, C. Yokoyama, M. Otsuki, S. Kihara, and I. Shimomura, "Adiponectin regulates vascular endothelial growth factor-c expression in macrophages via syk-erk pathway," *PloS one*, vol. 8, no. 2, p. e56071, 2013.

[129] W. W. Pang, E. A. Price, D. Sahoo, I. Beerman, W. J. Maloney, D. J. Rossi, S. L. Schrier, and I. L. Weissman, "Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age," *Proceedings of the National Academy of Sciences*, vol. 108, no. 50, pp. 20012–20017, 2011.

[130] S. M. Chambers, C. A. Shaw, C. Gatza, C. J. Fisk, L. A. Donehower, and M. A. Goodell, "Aging hematopoietic stem cells decline in function and exhibit epigenetic dysregulation," *PLoS biology*, vol. 5, no. 8, p. e201, 2007.

[131] S. M. Vanderwerf, J. Svahn, S. Olson, R. K. Rathbun, C. Harrington, J. Yates, W. Keeble, D. C. Anderson, P. Anur, N. F. Pereira, *et al.*, "Tlr8-dependent tnf-alpha overexpression in fanconi anemia group c cells," *Blood*, vol. 114, no. 26, pp. 5290–5298, 2009.

[132] T. Taniguchi and A. D. D'Andrea, "Molecular pathogenesis of fanconi anemia: recent progress," *Blood*, vol. 107, no. 11, pp. 4223–4233, 2006.

[133] A. Ridet, C. Guillouf, E. Duchaud, E. Cundari, M. Fiore, E. Moustacchi, and F. Rosselli, "Deregulated apoptosis is a hallmark of the fanconi anemia syndrome," *Cancer Research*, vol. 57, no. 9, pp. 1722–1730, 1997.

[134] R. Weksberg, M. Buchwald, P. Sargent, M. W. Thompson, and L. Siminovitch, "Specific cellular defects in patients with fanconi anemia," *Journal of cellular physiology*, vol. 101, no. 2, pp. 311–323, 1979.

[135] H. Joenje and K. J. Patel, "The emerging genetic and molecular basis of fanconi anaemia," *Nature Reviews Genetics*, vol. 2, no. 6, pp. 446–459, 2001.

[136] I. Medina, J. Carbonell, L. Pulido, S. C. Madeira, S. Goetz, A. Conesa, J. Tárraga, A. Pascual-Montano, R. Nogales-Cadenas, J. Santoyo, *et al.*, "Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling," *Nucleic acids research*, vol. 38, no. suppl 2, pp. W210–W213, 2010.

[137] F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo, "Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information," *Bioinformatics*, vol. 21, no. 13, pp. 2988–2993, 2005.

[138] H. Hurwitz, L. Fehrenbacher, W. Novotny, T. Cartwright, J. Hainsworth, W. Heim, J. Berlin, A. Baron, S. Griffing, E. Holmgren, *et al.*, "Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer," *New England journal of medicine*, vol. 350, no. 23, pp. 2335–2342, 2004.

[139] R. Fukuda, B. Kelly, and G. L. Semenza, "Vascular endothelial growth factor gene expression in colon cancer cells exposed to prostaglandin e2 is mediated by hypoxia-inducible factor 1," *Cancer research*, vol. 63, no. 9, pp. 2330–2334, 2003.

[140] X. Li, L. Shen, X. Shang, and W. Liu, "Subpathway analysis based on signaling-pathway impact analysis of signaling pathway," *PloS one*, vol. 10, no. 7, p. e0132813, 2015.

[141] T. Uehara, A. Ono, T. Maruyama, I. Kato, H. Yamada, Y. Ohno, and T. Urushidani, "The japanese toxicogenomics project: application of toxicogenomics," *Molecular nutrition and food research*, vol. 54, no. 2, pp. 218–227, 2010.

[142] M. Lukk, M. Kapushesky, J. Nikkila, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, and A. Brazma, "A global map of human gene expression," *Nature biotechnology*, vol. 28, no. 4, pp. 322–324, 2010.

[143] R. D. Hernansaiz-Ballesteros, F. Salavert, P. Sebastián-León, A. Alemán, I. Medina, and J. Dopazo, "Assessing the impact of mutations found in next generation sequencing data over human signaling pathways," *Nucleic acids research*, p. gk349, 2015.

[144] R. D. Hernansaiz-Ballesteros, "Variaciones alélicas en las rutas de señalización génica mediante estudios poblacionales y familiares," degree thesis, Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural, Universidad Politecnica de Valencia, Avenida de los Naranjos, s/n, 46022, Valencia, Spain, September 2012. 10.

[145] R. D. Hernansaiz-Ballesteros, "Implicaciones clínicas de la disfuncionalidad de rutas de señalización génica debido a las variaciones alélicas individuales," master thesis, Escuela técnica Superior de Ingeniería, Avinguda de la Universitat, s/n, 46100, Burjassot, Valencia, Spain, January 2014. 8.

[146] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, *et al.*, "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[147] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[148] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub, "A molecular signature of metastasis in primary solid tumors," *Nature genetics*, vol. 33, no. 1, pp. 49–54, 2003.

[149] J. P. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, *et al.*, "Repeatability of published microarray gene expression analyses," *Nature genetics*, vol. 41, no. 2, pp. 149–155, 2009.

[150] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: is there a unique set?," *Bioinformatics*, vol. 21, no. 2, pp. 171–178, 2005.

[151] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, pp. 5923–5928, 2006.

[152] B. Vogelstein, D. Lane, and A. J. Levine, "Surfing the p53 network," *Nature*, vol. 408, no. 6810, pp. 307–310, 2000.

[153] K. Mitra, A.-R. Carvunis, S. K. Ramesh, and T. Ideker, "Integrative approaches for finding modular structure in biological networks," *Nature Reviews Genetics*, vol. 14, no. 10, pp. 719–732, 2013.

[154] Z. Tu, C. Argmann, K. K. Wong, L. J. Mitnaul, S. Edwards, I. C. Sach, J. Zhu, and E. E. Schadt, "Integrating sirna and protein–protein interaction data to identify an expanded insulin signaling network," *Genome research*, vol. 19, no. 6, pp. 1057–1067, 2009.

[155] X. Ma, H. Lee, L. Wang, and F. Sun, "Cgi: a new approach for prioritizing genes by combining gene expression and protein–protein interaction data," *Bioinformatics*, vol. 23, no. 2, pp. 215–221, 2007.

[156] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular systems biology*, vol. 3, no. 1, 2007.

[157] Y.-Q. Qiu, S. Zhang, X.-S. Zhang, and L. Chen, "Detecting disease associated modules and prioritizing active genes based on high throughput data," *BMC bioinformatics*, vol. 11, no. 1, p. 26, 2010.

[158] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nature biotechnology*, vol. 27, no. 2, pp. 199–204, 2009.

[159] G. Ciriello, E. Cerami, C. Sander, and N. Schultz, "Mutual exclusivity analysis identifies oncogenic network modules," *Genome research*, vol. 22, no. 2, pp. 398–406, 2012.

[160] S. A. Bapat, A. Krishnan, A. D. Ghanate, A. P. Kusumbe, and R. S. Kalra, "Gene expression: protein interaction systems network modeling identifies transformation-associated molecules and pathways in ovarian cancer," *Cancer research*, vol. 70, no. 12, pp. 4809–4819, 2010.

[161] K. X. Zhang and B. F. Ouellette, "Caerus: predicting cancer outcomes using relationship between protein structural information, protein networks, gene expression data, and mutation data," *PLoS computational biology*, vol. 7, no. 3, p. e1001114, 2011.

[162] J. Ahn, Y. Yoon, C. Park, E. Shin, and S. Park, "Integrative gene network construction for predicting a set of complementary prostate cancer genes," *Bioinformatics*, vol. 27, no. 13, pp. 1846–1853, 2011.

[163] P. Dao, K. Wang, C. Collins, M. Ester, A. Lapuk, and S. C. Sahinalp, "Optimally discriminative subnetwork markers predict response to chemotherapy," *Bioinformatics*, vol. 27, no. 13, pp. i205–i213, 2011.

[164] Z. Wu, X.-M. Zhao, and L. Chen, "A systems biology approach to identify effective cocktail drugs," *BMC systems biology*, vol. 4, no. Suppl 2, p. S7, 2010.

[165] P. Romanski, L. Kotthoff, and M. L. Kotthoff, "Package 'fselector'," 2013.

[166] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[167] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, *et al.*, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, no. 7391, pp. 570–575, 2012.

[168] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.

[169] A. Karatzoglou, D. Meyer, and K. Hornik, "Support vector machines in r," 2005.

[170] F. Eduati, L. M. Mangravite, T. Wang, H. Tang, J. C. Bare, R. Huang, T. Norman, M. Kellen, M. P. Menden, J. Yang, *et al.*, "Prediction of human population responses to toxic compounds by a collaborative competition," *Nature Biotechnology*, 2015.

[171] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, "The sequence of the human genome," *science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[172] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.-J. Chen, V. Makhijani, G. T. Roth, *et al.*, "The complete

genome of an individual by massively parallel dna sequencing," *nature*, vol. 452, no. 7189, pp. 872–876, 2008.

[173] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol, "Human genetic variation and its contribution to complex traits," *Nature Reviews Genetics*, vol. 10, no. 4, pp. 241–251, 2009.

[174] K. Wang, M. Li, and H. Hakonarson, "Annovar: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic acids research*, vol. 38, no. 16, pp. e164–e164, 2010.

[175] P. C. Ng and S. Henikoff, "Predicting deleterious amino acid substitutions," *Genome research*, vol. 11, no. 5, pp. 863–874, 2001.

[176] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature methods*, vol. 7, no. 4, pp. 248–249, 2010.

[177] . G. P. Consortium *et al.*, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[178] P. Warren, D. Taylor, P. G. Martini, J. Jackson, and J. Bienkowska, "Panp-a new method of gene detection on oligonucleotide expression arrays," in *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, pp. 108–115, IEEE, 2007.

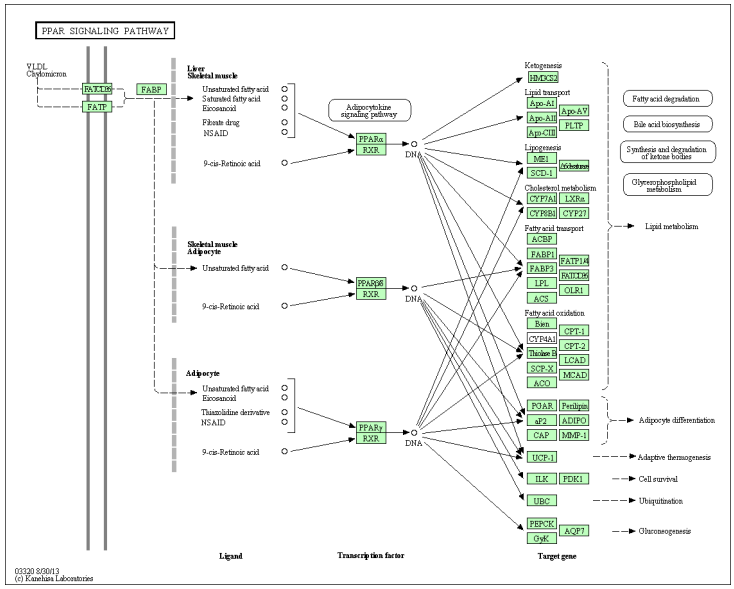
[179] S. Myles, M. Somel, K. Tang, J. Kelso, and M. Stoneking, "Identifying genes underlying skin pigmentation differences among human populations," *Human genetics*, vol. 120, no. 5, pp. 613–621, 2007.

Chapter

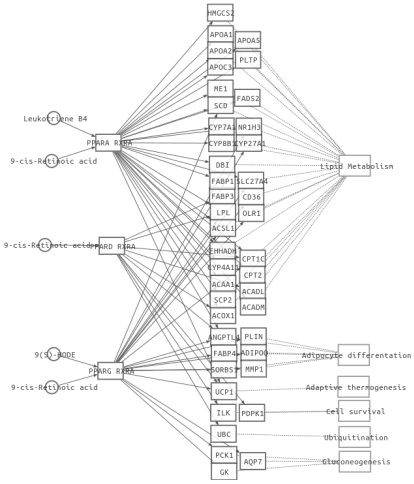
A Modeled KEGG pathways

In this chapter modeled pathways for *Homo sapiens* are represented joined with its representation in KEGG database. Pathways of other species are quite similar, and can be observed in PATHiWAYS web-tool [89].

Pathway name	Reference
PPAR signaling pathway	Figure A.1
ErbB signaling pathway	Figure A.2
Calcium Signaling pathway	Figure A.3
Cytokine-cytokine receptor interaction	Figure A.4
Chemokine signaling pathway	Figure A.5
Neuroactive ligand-receptor interaction	Figure A.6
p53 signaling pathway	Figure A.7
mTOR signaling pathway	Figure A.8
Apoptosis	Figure A.9
Wnt signaling pathway	Figure A.10
Notch Signaling pathway	Figure A.11
Hedgehog signaling pathway	Figure A.12
VEGF signaling pathway	Figure A.13
ECM-receptor interaction	Figure A.14
Cell adhesion molecules (CAMs)	Figure A.15
Tight junction	Figure A.16
Gap junction	Figure A.17
Antigen processing and presentation	Figure A.18
Toll-like receptor signaling pathway	Figure A.19
Jak-STAT signaling pathway	Figure A.20
T cell receptor signaling pathway	Figure A.21
B cell receptor signaling pathway	Figure A.22
Fc epsilon RI signaling pathway	Figure A.23
Insulin signaling pathway	Figure A.24
GnRH signaling pathway	Figure A.25
Melanogenesis	Figure A.26
Adipocytokine signaling pathway	Figure A.27

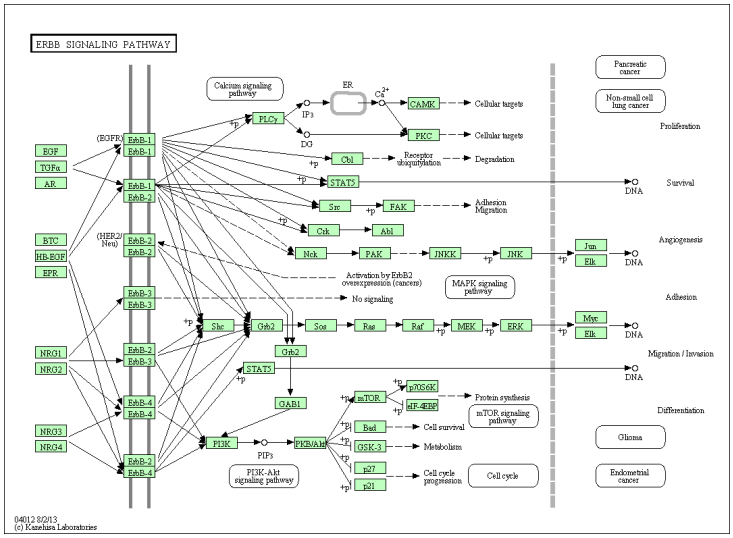


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa03320`)

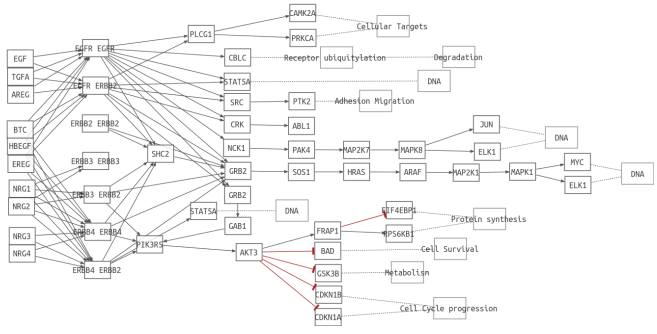


(b) Modeled according KGML information

Figure A.1: PPAR signaling pathway (hsa03320)

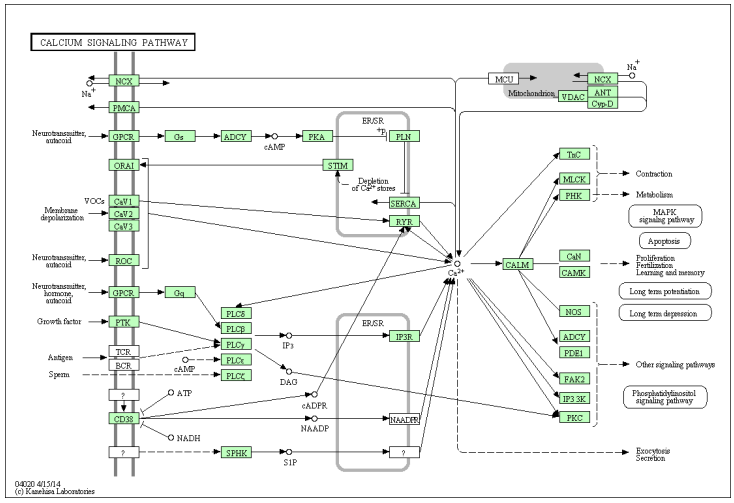


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04012`)

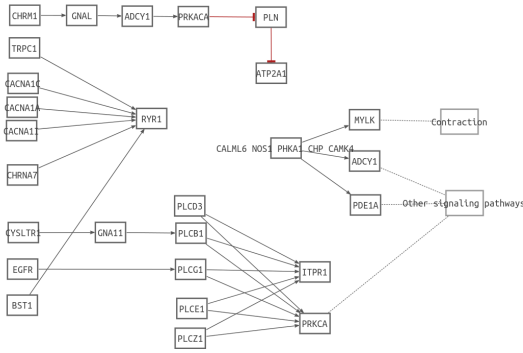


(b) Modeled according KGML information

Figure A.2: ErbB Signalig pathway (hsa04012)

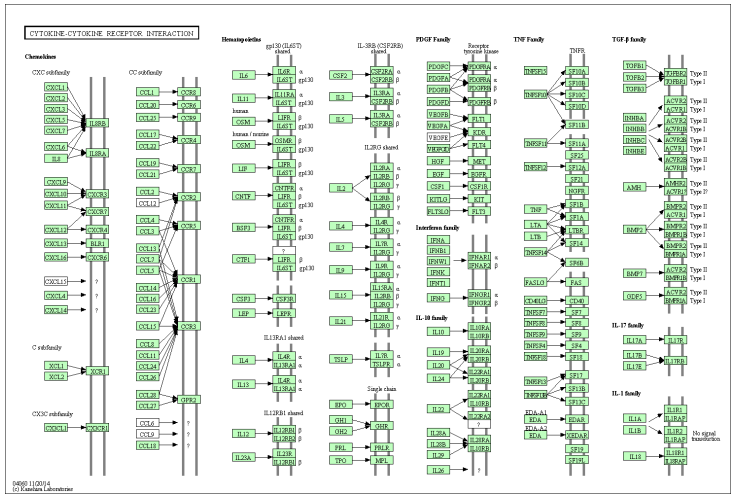


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04020`)

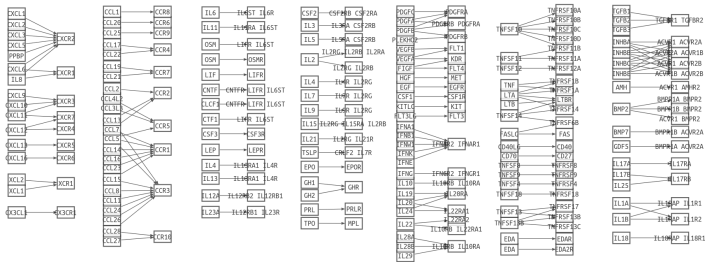


(b) Modeled according KGML information

Figure A.3: Calcium Signaling pathway

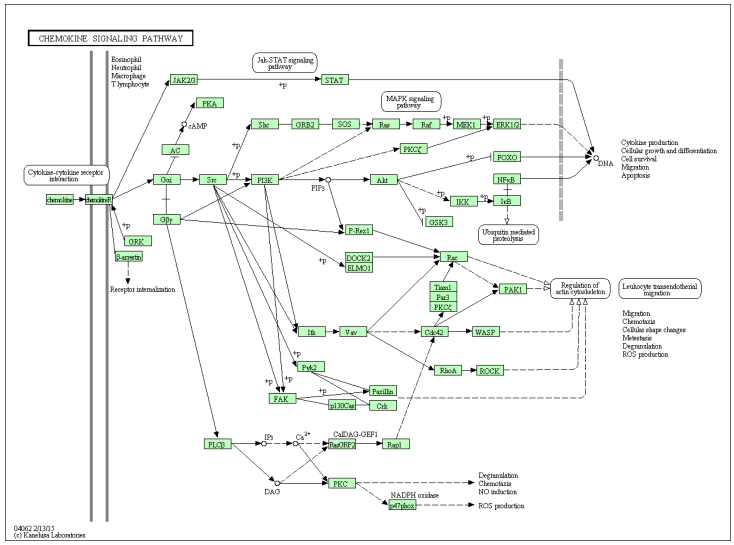


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04060`)

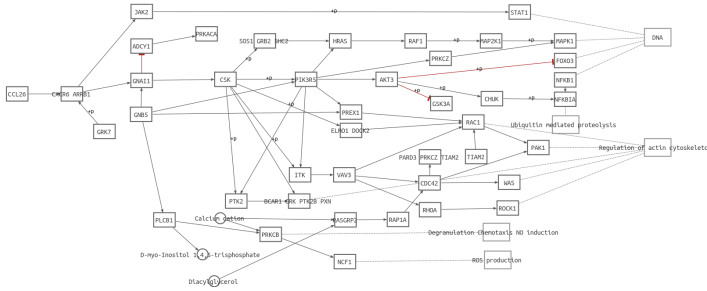


(b) Modeled according KGML information

Figure A.4: Cytokine-cytokine receptor interaction (hsa04060)

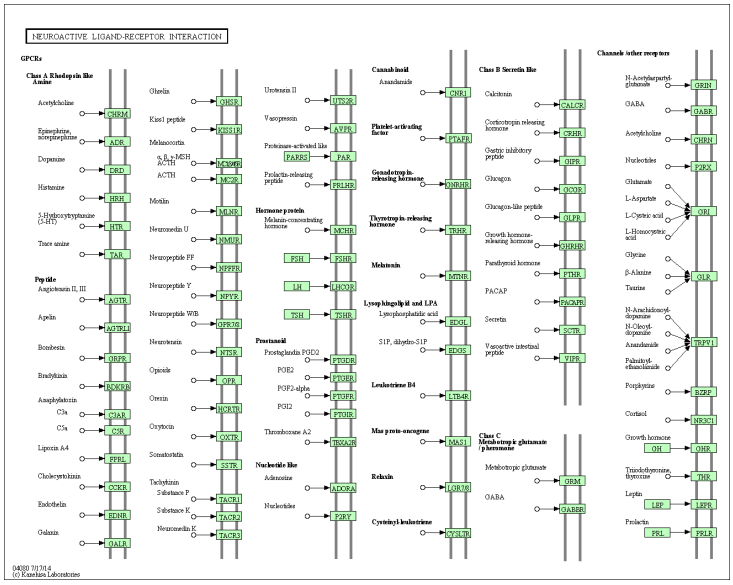


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04062`)

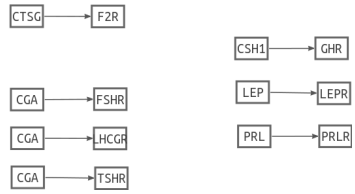


(b) Modeled according KGML information

Figure A.5: Chemokine signaling pathway (hsa04062)

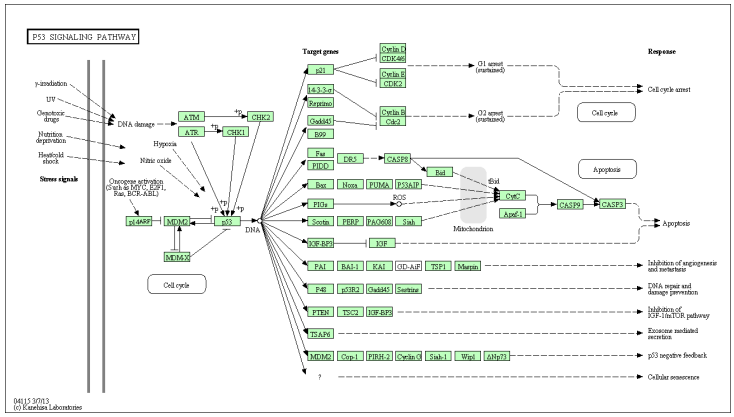


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04080`)

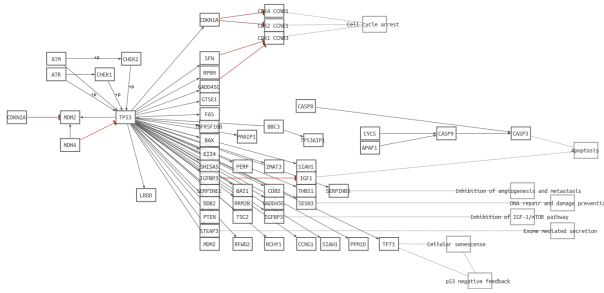


(b) Modeled according KGML information

Figure A.6: Neuroactive ligand-receptor interaction (hsa04080)

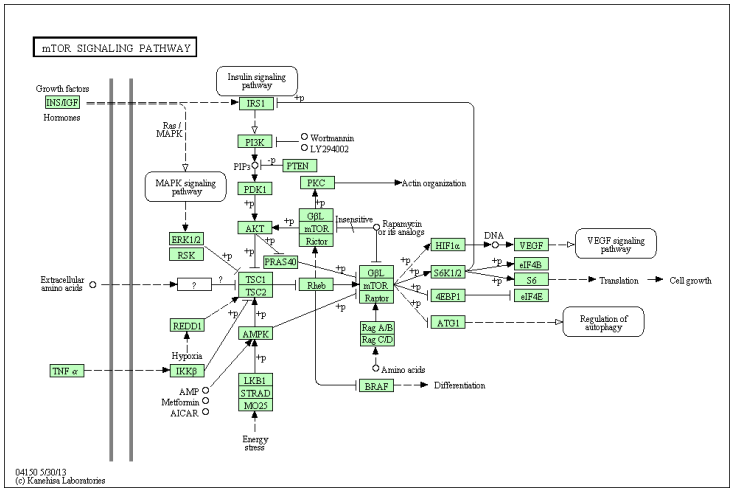


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04115`)

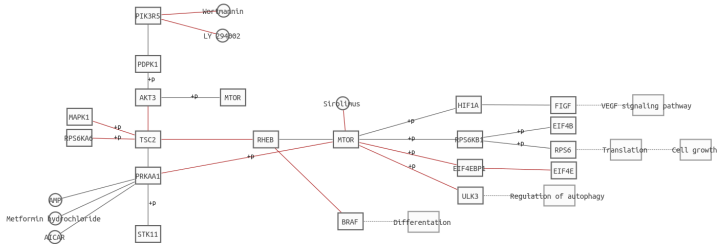


(b) Modeled according KGML information

Figure A.7: p53 signaling pathway (hsa04115)

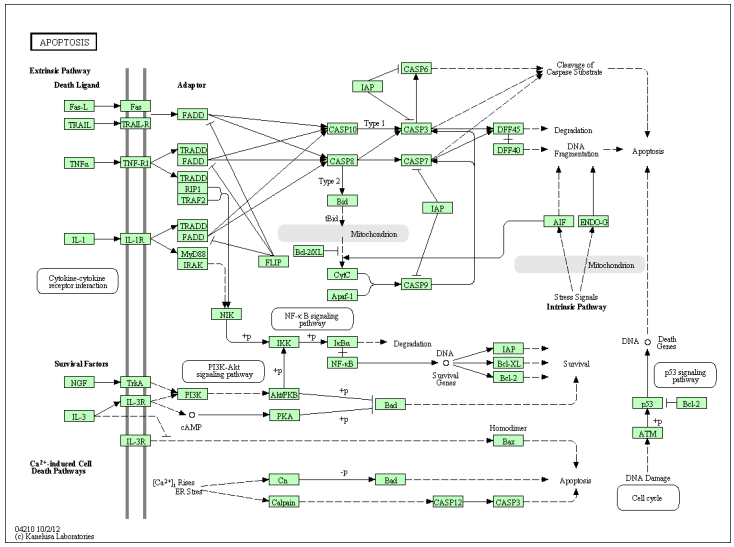


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04150`)

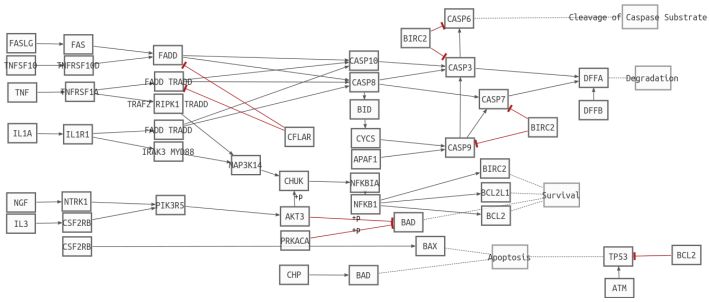


(b) Modeled according KGML information

Figure A.8: mTOR signaling pathway (hsa04150)

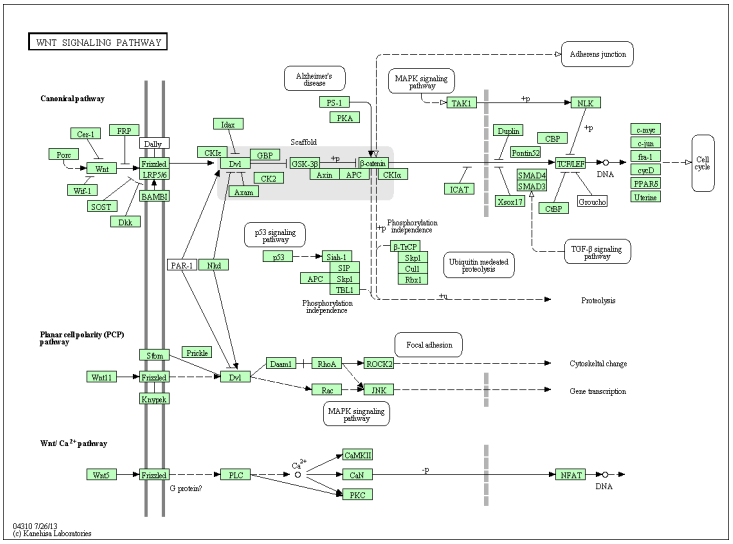


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04210`)

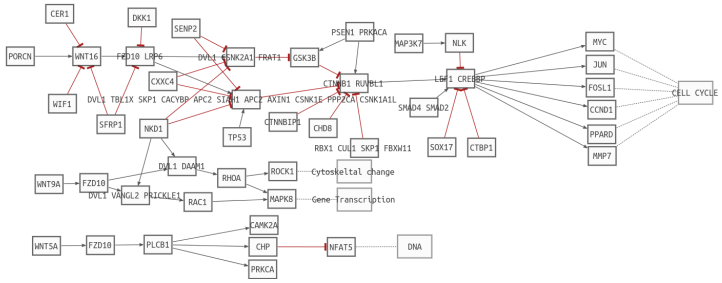


(b) Modeled according KGML information

Figure A.9: Apoptosis (hsa04210)

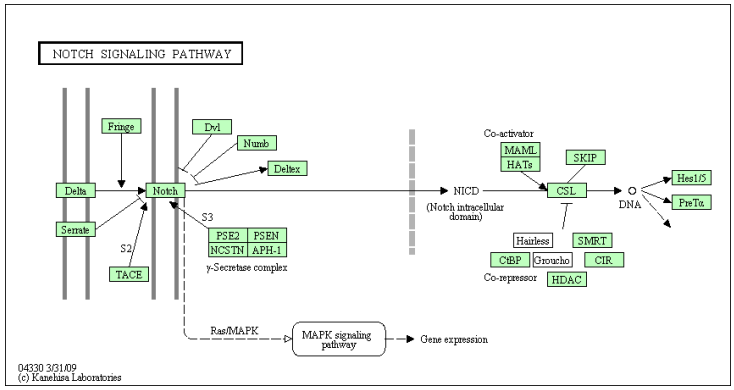


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04310`)

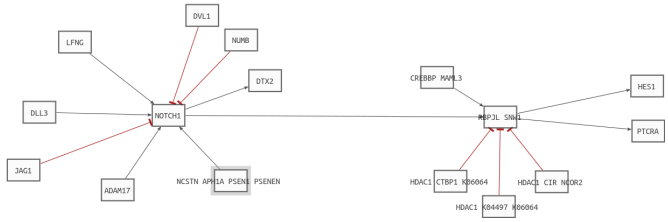


(b) Modeled according KGML information

Figure A.10: Wnt signaling pathway (hsa04310)

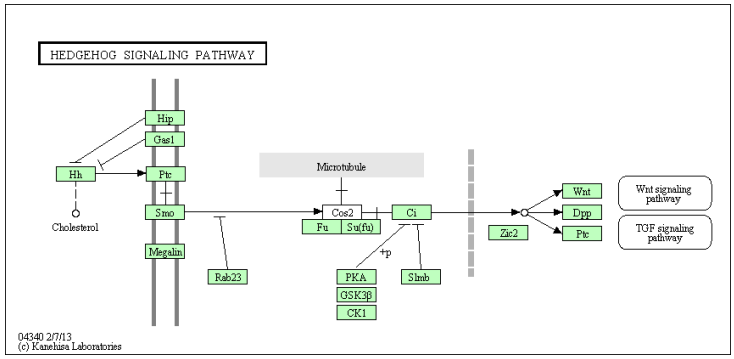


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04330`)

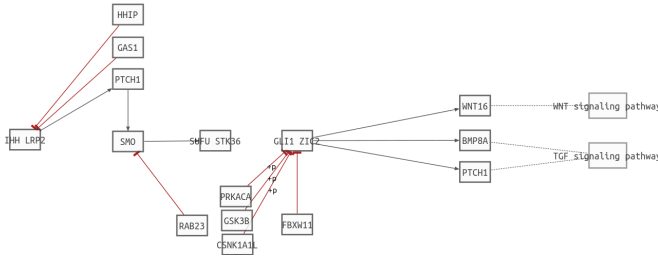


(b) Modeled according KGML information

Figure A.11: Notch Signaling pathway (hsa04330)

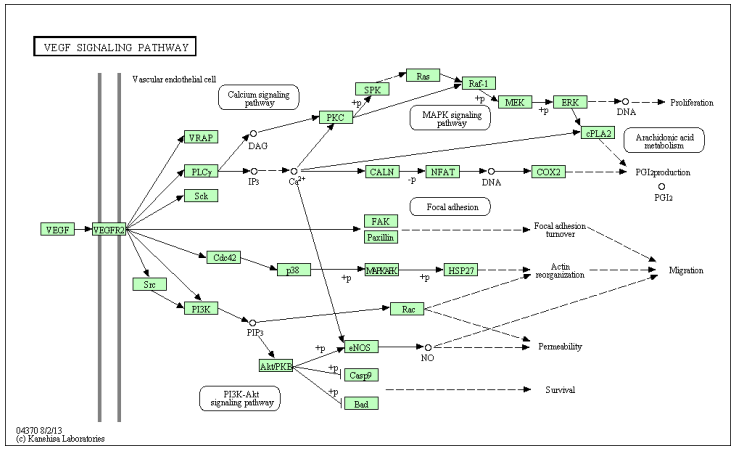


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04340`)

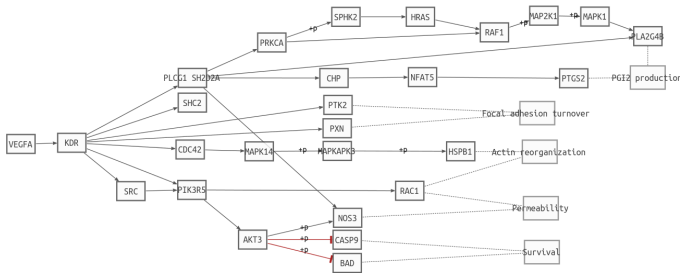


(b) Modeled according KGML information

Figure A.12: Hedgehog signaling pathway (hsa04340)

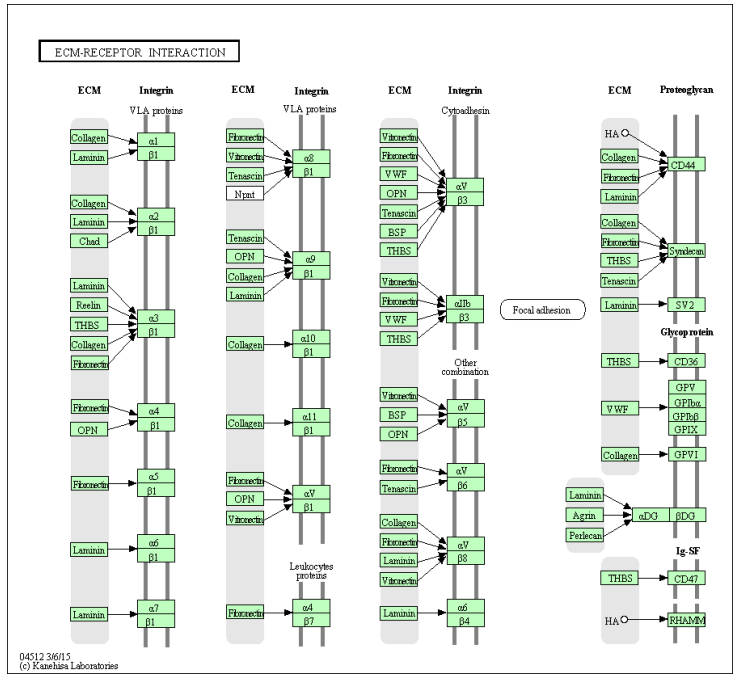


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04370`)

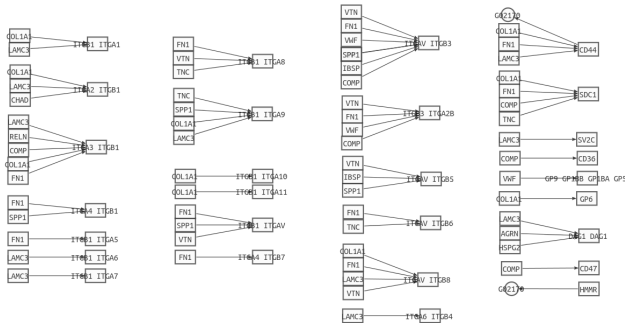


(b) Modeled according KGML information

Figure A.13: VEGF signaling pathway (hsa04370)

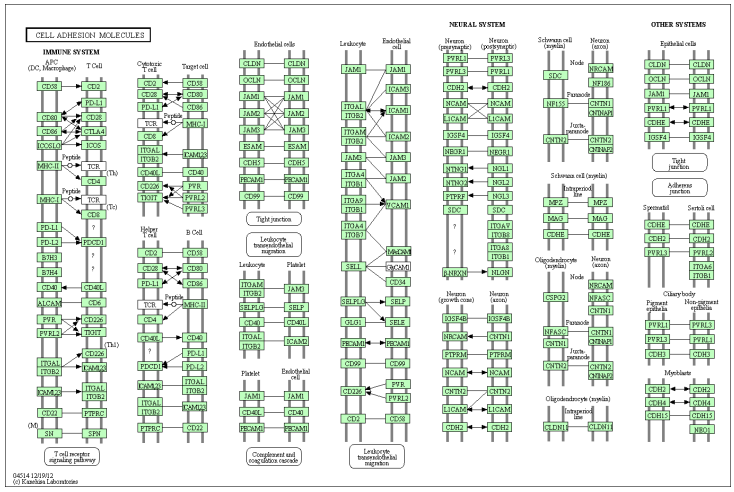


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04512`)

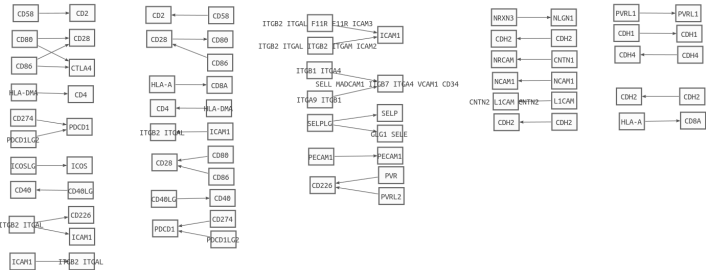


(b) Modeled according KGML information

Figure A.14: ECM-receptor interaction (hsa04512)

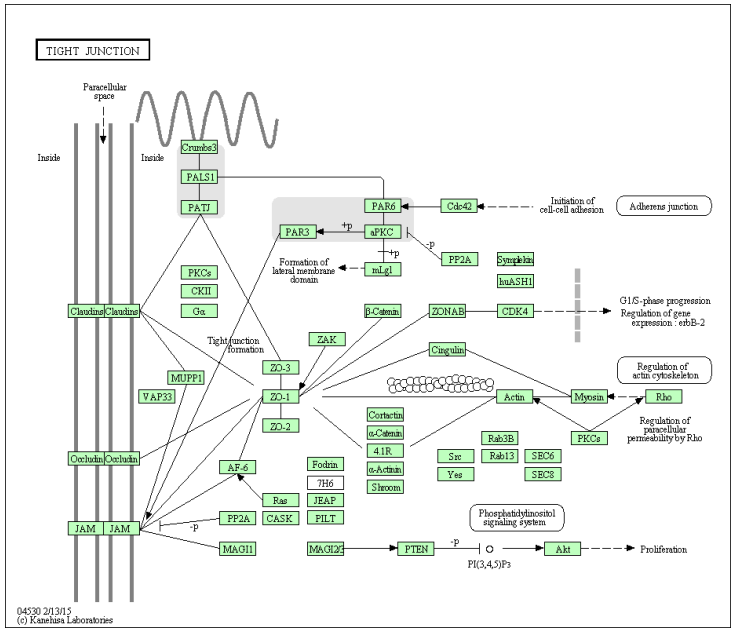


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04514`)

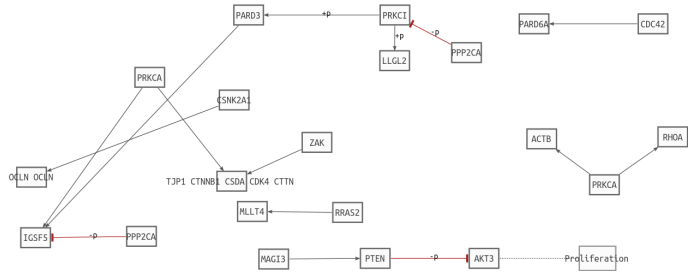


(b) Modeled according KGML information

Figure A.15: Cell adhesion molecules (CAMs) (hsa04514)

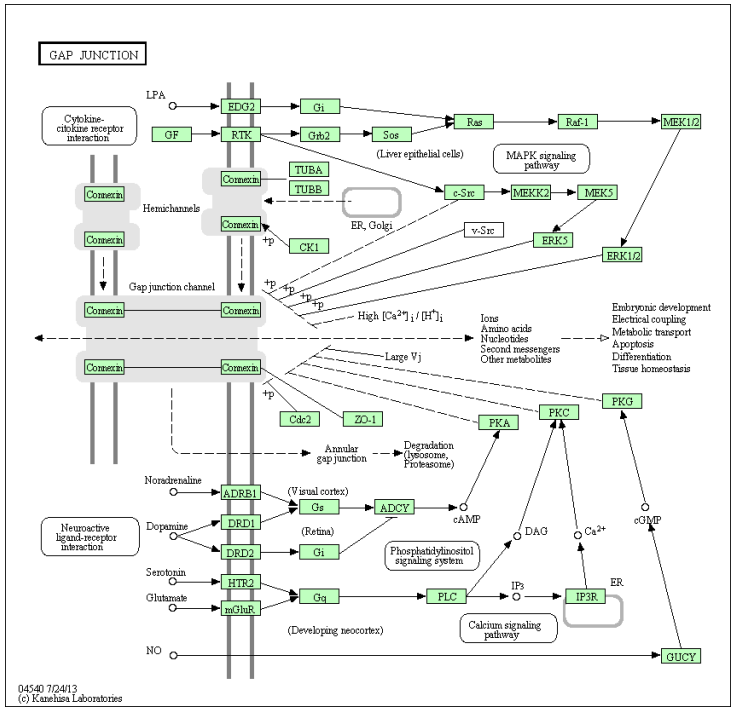


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04530`)

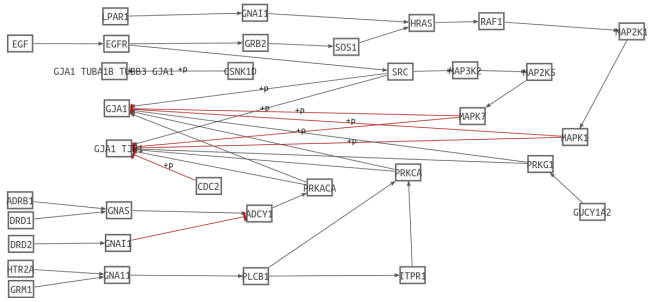


(b) Modeled according KGML information

Figure A.16: Tight junction (hsa04530)

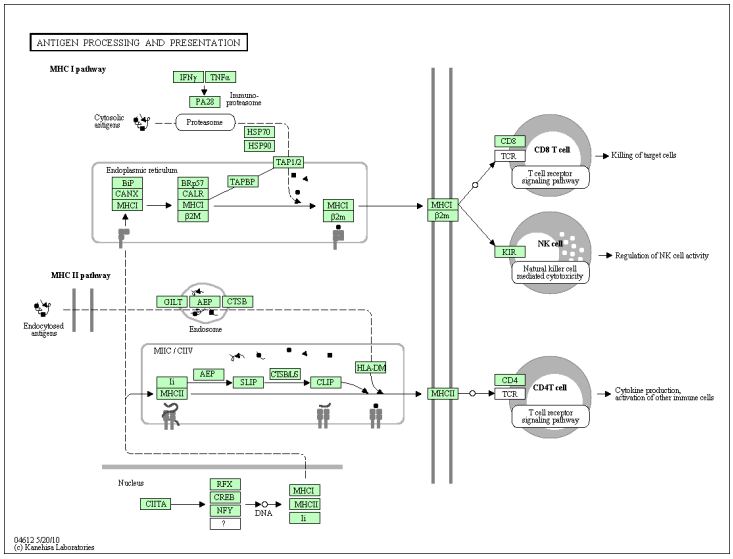


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04540`)

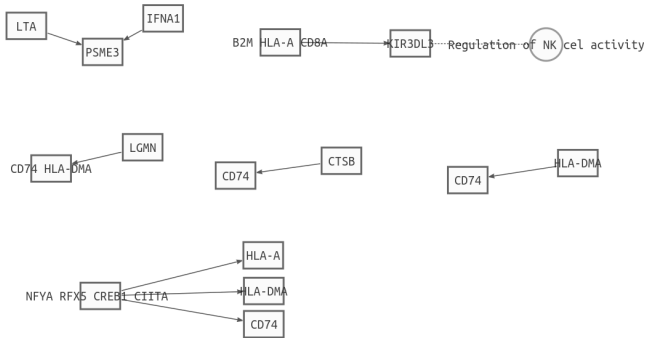


(b) Modeled according KGML information

Figure A.17: Gap junction (hsa04540)

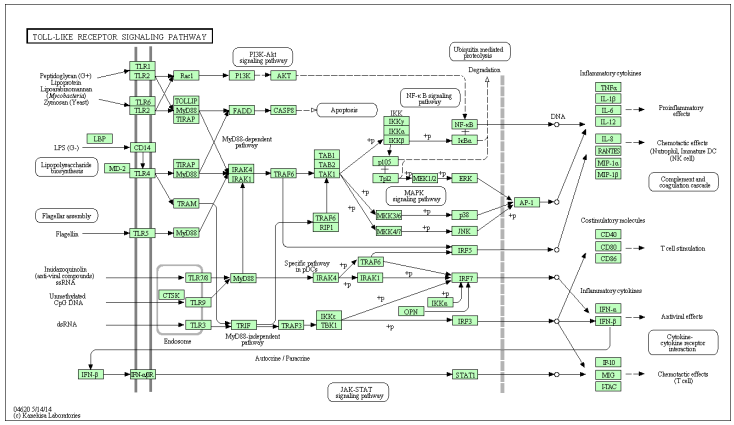


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04612`)

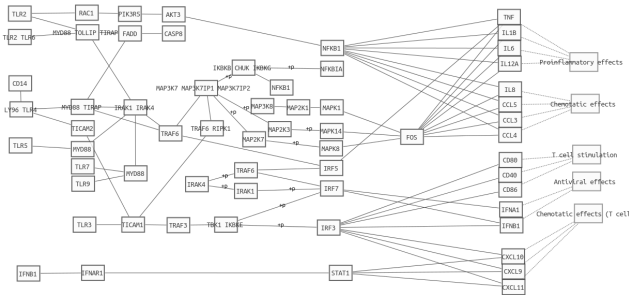


(b) Modeled according KGML information

Figure A.18: Antigen processing and presentation (hsa04612)

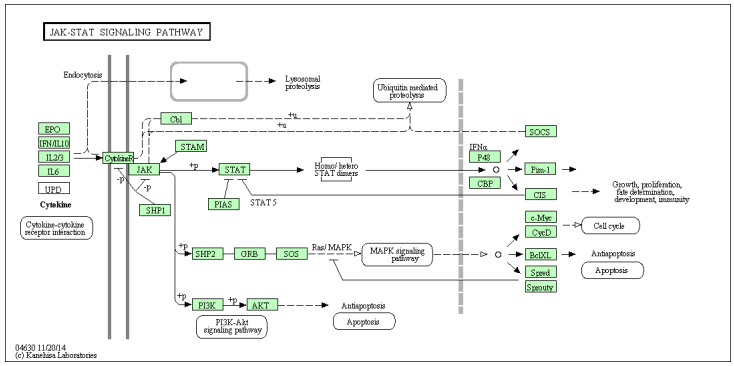


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04620`)

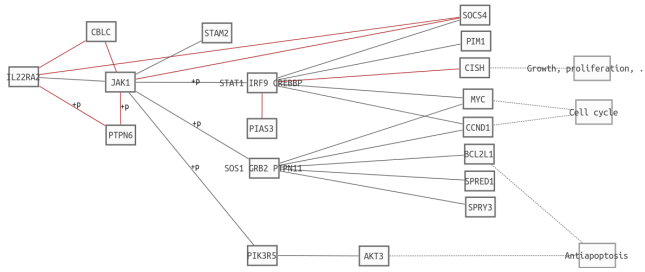


(b) Modeled according KGML information

Figure A.19: Toll-like receptor signaling pathway (hsa04620)

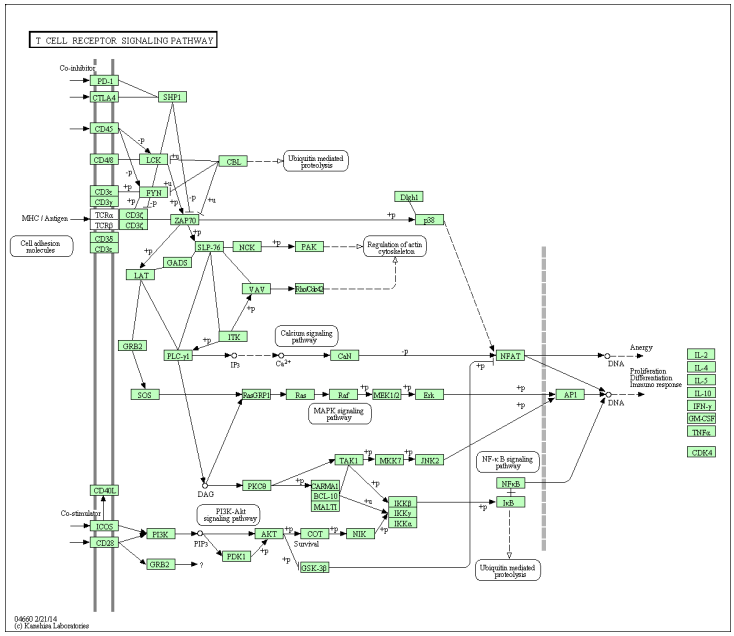


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04630`)

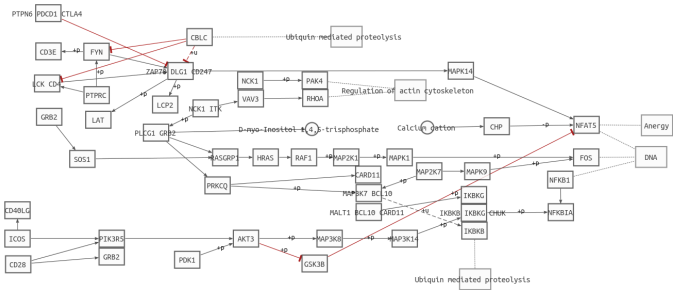


(b) Modeled according KGML information

Figure A.20: Jak-STAT signaling pathway (hsa04630)

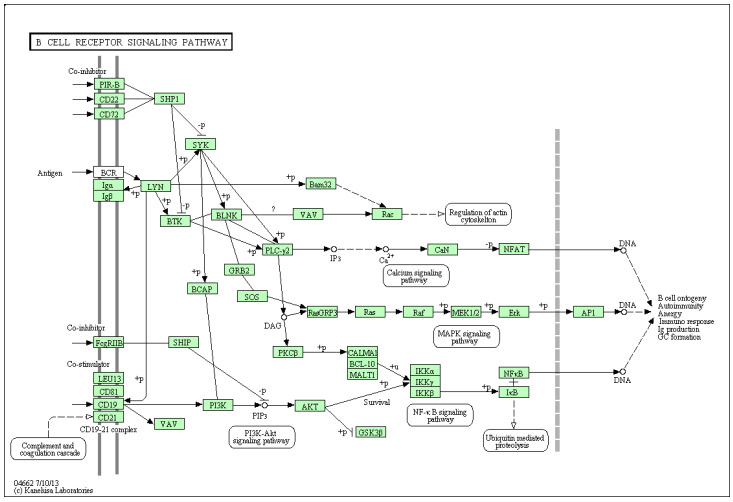


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04660`)

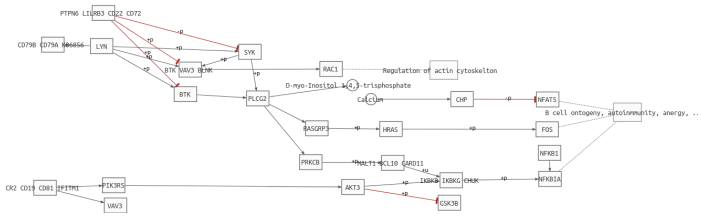


(b) Modeled according KGML information

Figure A.21: T cell receptor signaling pathway (hsa04660)

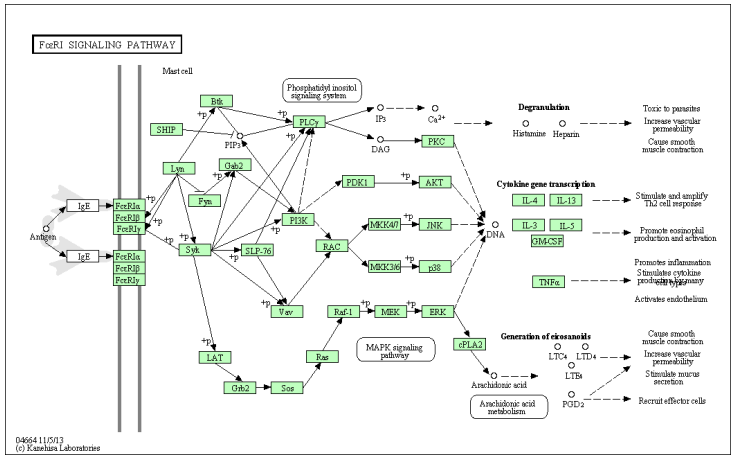


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04662`)

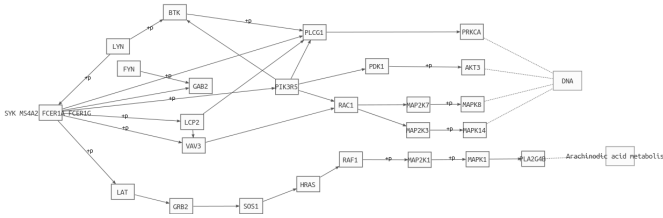


(b) Modeled according KGML information

Figure A.22: B cell receptor signaling pathway (hsa04662)

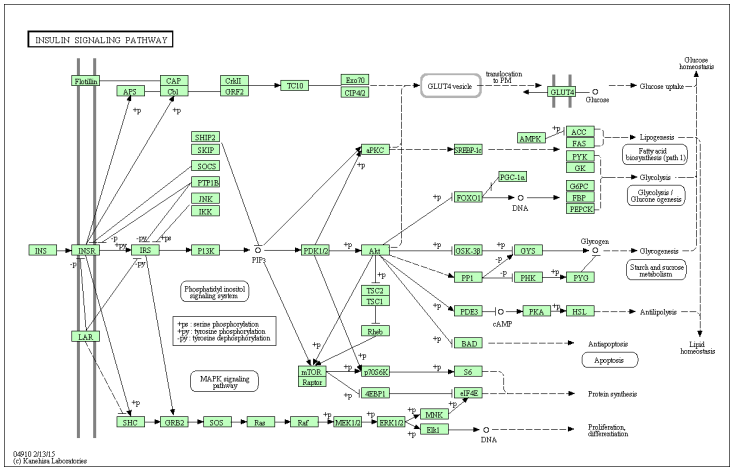


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04664`)

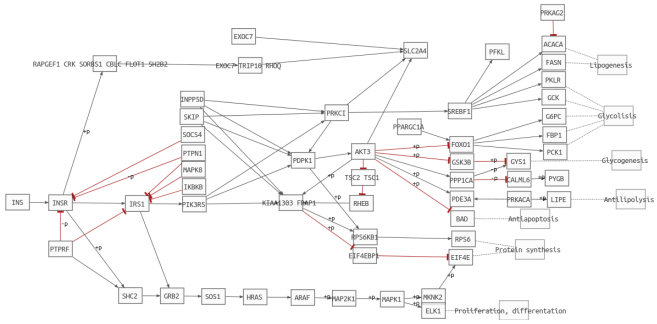


(b) Modeled according KGML information

Figure A.23: Fc epsilon RI signaling pathway (hsa04664)

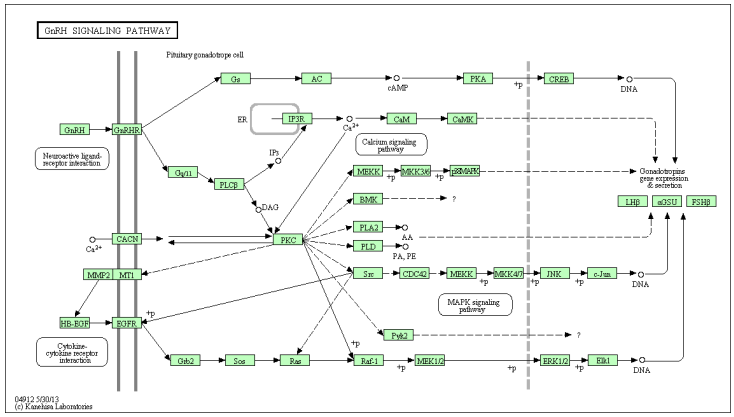


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04910`)

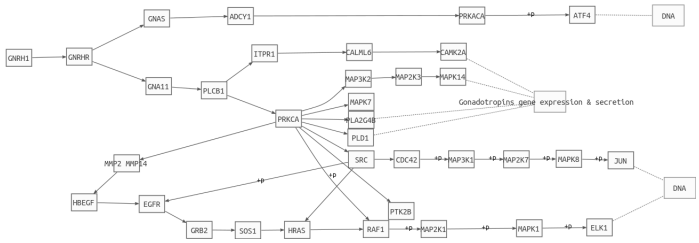


(b) Modeled according KGML information

Figure A.24: Insulin signaling pathway (hsa04910)

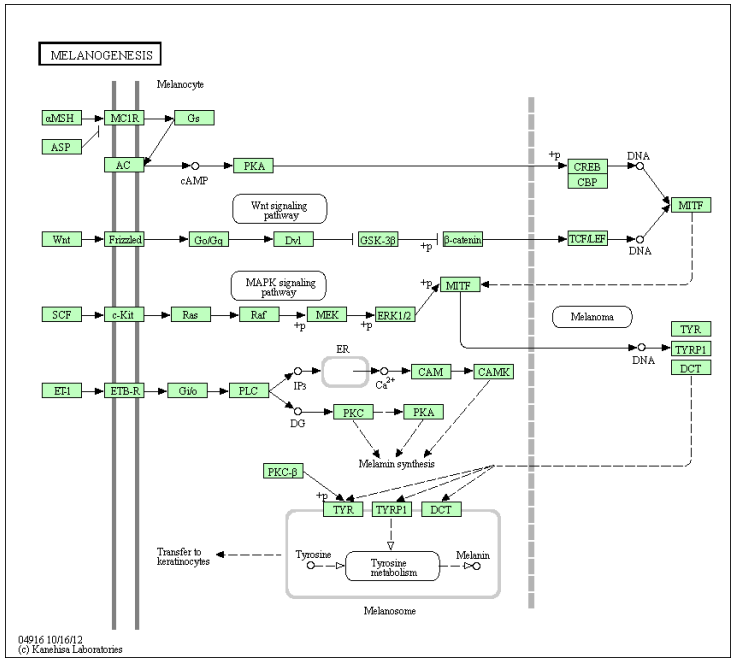


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04912`)

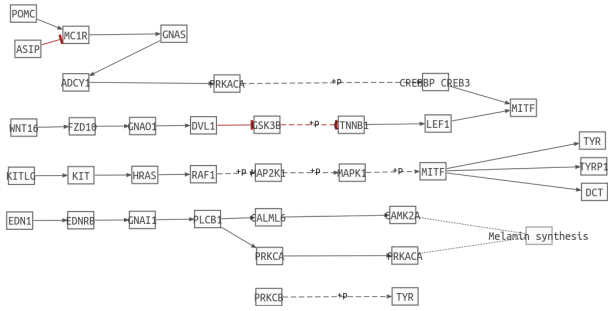


(b) Modeled according KGML information

Figure A.25: GnRH signaling pathway (hsa04912)

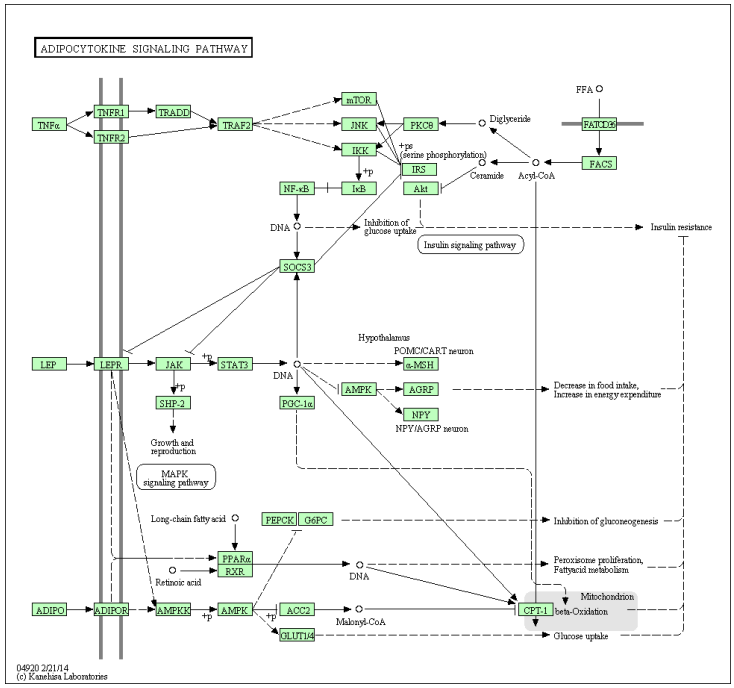


(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04916`)

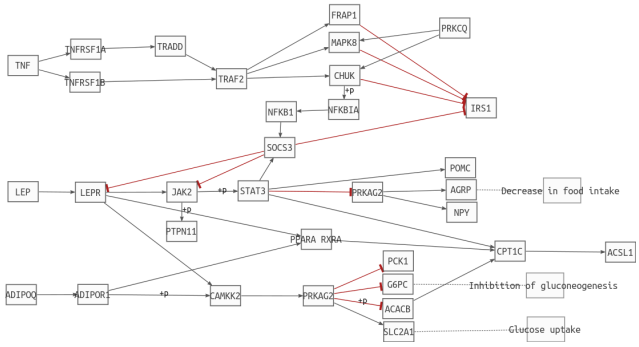


(b) Modeled according KGML information

Figure A.26: Melanogenesis (hsa04916)



(a) KEGG representation (`http://www.genome.jp/kegg-bin/show_pathway?hsa04920`)



(b) Modeled according KGML information

Figure A.27: Adipocytokine signaling pathway (hsa04920)

Chapter

B Complete results of CRC data analysis

Summary results of the analysis of this dataset using Wilcoxon testing can be consulted in table B.1. Here you can find pathways not included in results section, tables as well as representations.

Pathway name	significant/total	UP	DOWN
PPAR signaling pathway	22/106	3	19
ErbB signaling pathway	13/139	11	2
Calcium signaling pathway	4/20	2	2
Cytokine-cytokine receptor interaction	21/179	19	2
Chemokine signalin pathway	5/43	3	2
Neuroactive ligand-receptor interaction	0/7	0	0
p53 signaling pathway	0/150	0	0
mTOR signaling pathway	0/36	0	0
Apoptosis	0/28	0	0
WNT signaling pathway	12/37	6	6
Notch signaling pathway	0/14	0	0
Hedgehog signaling pathway	0/4	0	0
VEGF signaling pathway	2/10	2	0
ECM-receptorinteraction	10/65	10	0
Cell adhesion molecules	9/43	6	3
Tight junction	2/11	2	0
Gap junction	4/11	4	0
Antigen procesing and presentation	0/6	0	0
Toll-like receptor signaling pathway	0/103	0	0
Jak-STAT signaling pathway	7/7	7	0
T cell receptor signaling pathway	6/24	5	1
B cell receptor signaling pathway	0/10	0	0
Fc epsilon RI signaling pathway	0/7	0	0
nsulin signaling pathway	1/54	1	0
GnRH signaling pathway	0/9	0	0
Melanogenesis	1/8	1	0
Adipocytokine signaling pathway	2/31	0	2

Table B.1: Summary results of colorectal cancer experiment (GSE4107)

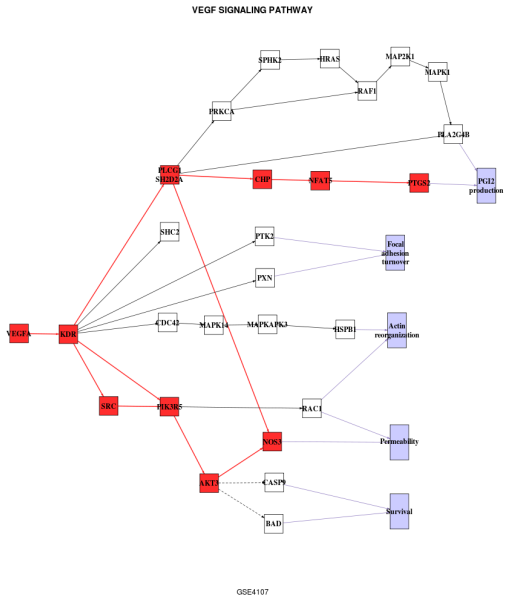


Figure B.1: VEGF signaling pathway results in CRC study

VEGF SIGNALING PATHWAY (hsa04370)			
Subpathway	p-value	FDR p.value	UP/DOWN
VEGFA-PTGS2	0.0002	0.0023	UP
VEGFA-NOS3	0.0074	0.0372	UP

Table B.2: VEGF results in colorectal cancer study

JAK-STAT SIGNALING PATHWAY (hsa04630)			
Subpathway	p-value	FDR p.value	UP/DOWN
STAM2-MYC	0.0057	0.0130	UP
STAM2-AKT3	0.0074	0.0130	UP
STAM2-PIM1	0.0031	0.0130	UP
STAM2-CCND1	0.0042	0.0130	UP
STAM2-BCL2L1	0.0310	0.0381	UP
STAM2-SPRED1	0.0381	0.0381	UP
STAM2-SPRY3	0.0381	0.0381	UP

Table B.3: Jak-STAT signaling pathway results in colorectal cancer study

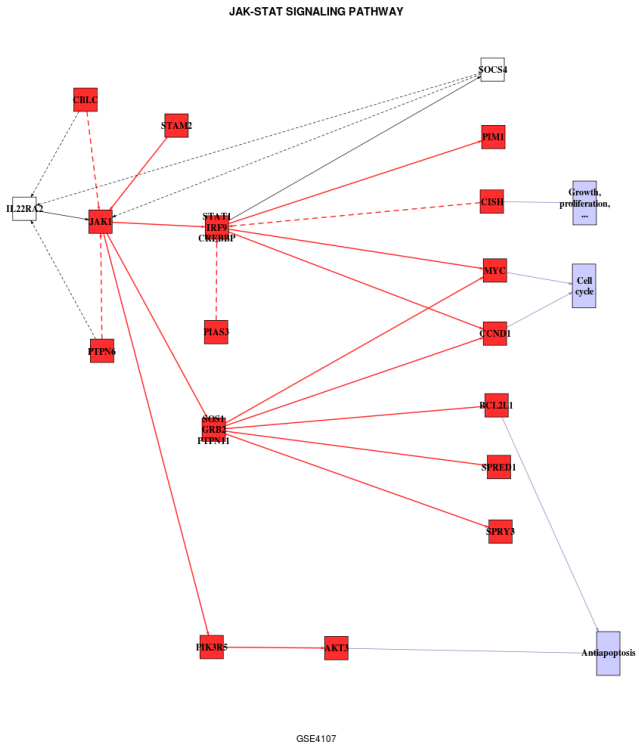


Figure B.2: Jak-STAT signaling pathway results in CRC study

WNT SIGNALING PATHWAY (hsa04310)			
Subpathway	p-value	FDR p.value	UP/DOWN
PORCN-MYC	0.0042	0.019	DOWN
PORCN-JUN	0.0042	0.019	DOWN
PORCN-FOSL1	0.0016	0.019	DOWN
PORCN-CCND1	0.0057	0.019	DOWN
PORCN-PPARD	0.0031	0.019	DOWN
MAP3K7-MYC	0.0042	0.019	UP
MAP3K7-JUN	0.0031	0.019	UP
MAP3K7-CCND1	0.0057	0.019	UP
MAP3K7-PPARD	0.0057	0.019	UP
MAP3K7-MMP7	0.0023	0.019	UP
WNT9A-MAPK8	0.0031	0.019	UP
PORCN-MMP7	0.0159	0.049	DOWN

Table B.4: WNT signaling pathway results in CRC study

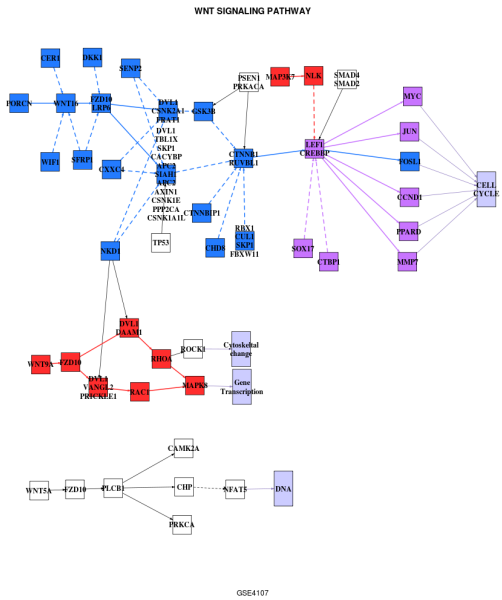


Figure B.3: WNT signaling pathway results in CRC study

PPAR SIGNALING PATHWAY (hsa03320)			
Subpathway	p-value	FDR p.value	UP/DOWN
9-cis-Retinoic_acid-CPT1C	0.0000	0.0017	DOWN
9-cis-Retinoic_acid-CPT1C	0.0000	0.0017	DOWN
9(S)-HODE-CPT1C	0.0000	0.0017	DOWN
9-cis-Retinoic_acid-CPT1C	0.0002	0.0048	DOWN
LeukotrieneB4-CPT1C	0.0002	0.0048	DOWN
9-cis-Retinoic_acid-LPL	0.0074	0.0438	UP
9-cis-Retinoic_acid-GK	0.0074	0.0438	DOWN
9-cis-Retinoic_acid-SCD	0.0057	0.0438	DOWN
9-cis-Retinoic_acid-MMP1	0.0057	0.0438	DOWN
9(S)-HODE-GK	0.0074	0.0438	DOWN
9(S)-HODE-SCD	0.0057	0.0438	DOWN
9(S)-HODE-MMP1	0.0057	0.0438	DOWN
9-cis-Retinoic_acid-PLTP	0.0074	0.0438	UP
9-cis-Retinoic_acid-CPT2	0.0057	0.0438	DOWN
9-cis-Retinoic_acid-EHHADH	0.0074	0.0438	DOWN
LeukotrieneB4-PLTP	0.0074	0.0438	UP
LeukotrieneB4-CPT2	0.0057	0.0438	DOWN
LeukotrieneB4-EHHADH	0.0074	0.0438	DOWN
9-cis-Retinoic_acid-APOA2	0.0097	0.0467	DOWN
9-cis-Retinoic_acid-ACAA1	0.0097	0.0467	DOWN
LeukotrieneB4-APOA2	0.0097	0.0467	DOWN
LeukotrieneB4-ACAA1	0.0097	0.0467	DOWN

Table B.5: PPAR signaling pathway results in CRC study

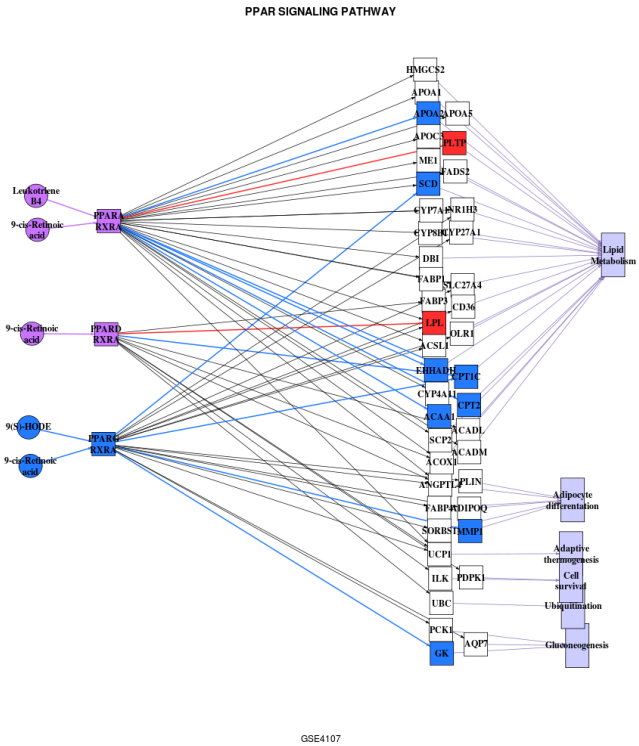


Figure B.4: PPAR signaling pathway results in CRC study

ERBB SIGNALING PATHWAY (hsa04012)			
Subpathway	p-value	FDR p.value	UP/DOWN
HBEGF-STAT5A	0.0000	0.0033	UP
HBEGF-CAMK2A	0.0000	0.0033	UP
HBEGF-STAT5A	0.0001	0.0039	UP
HBEGF-ABL1	0.0004	0.0124	UP
HBEGF-JUN	0.0005	0.0124	UP
HBEGF-PRKCA	0.0005	0.0124	UP
TGFA-CBLC	0.0008	0.0158	DOWN
HBEGF-ELK1	0.0012	0.0178	UP
TGFA-ABL1	0.0012	0.0178	DOWN
HBEGF-CBLC	0.0016	0.0228	UP
HBEGF-BAD	0.0042	0.0453	UP
HBEGF-CDKN1B	0.0042	0.0453	UP
HBEGF-GSK3B	0.0042	0.0453	UP

Table B.6: ERBB signaling pathway results in CRC study

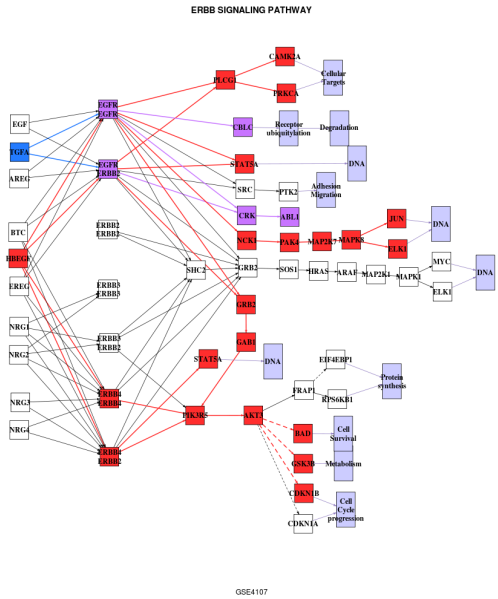


Figure B.5: ERBB signaling pathway results in CRC study

CALCIUM SIGNALING PATHWAY (hsa04020)			
Subpathway	p-value	FDR p.value	UP/DOWN
CHRM1-ATP2A1	0.0003	0.0053	DOWN
PLCE1-ITPR1	0.0023	0.0229	DOWN
CALML6 NOS1 PHKA1 CHP CAMK4-PDE1A	0.0057	0.0377	UP
CALML6 NOS1 PHKA1 CHP CAMK4-MYLK	0.0097	0.0485	UP

Table B.7: Calcium signaling pathway results in CRC study

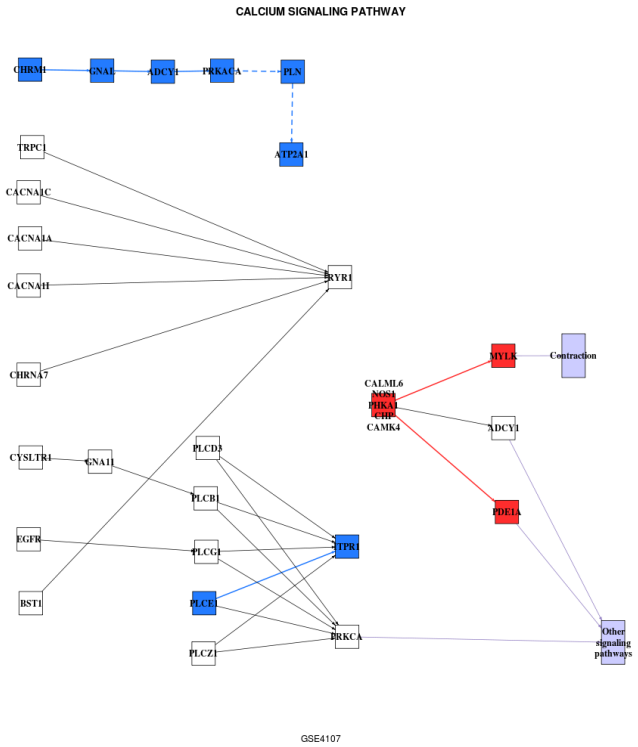


Figure B.6: Calcium signaling pathway results in CRC study

CYTOKINE-CYTOKINE RECEPTOR INTERACTION (hsa04060)			
Subpathway	p-value	FDR p.value	UP/DOWN
PDGFB-PDGFRB PDGFRA	0.0000	0.0011	UP
PDGFB-PDGFRB	0.0000	0.0011	UP
PDGFB-PDGFRA	0.0000	0.0014	UP
CXCL12-CXCR4	0.0001	0.0037	UP
TNFSF10-TNFRSF10D	0.0002	0.0067	DOWN
CXCL12-CXCR7	0.0002	0.0067	UP
INHBB-ACVR1 ACVR2A	0.0016	0.0331	UP
INHBB-ACVR2A ACVR1B	0.0016	0.0331	UP
CX3CL1-CX3CR1	0.0023	0.0331	UP
CCL7-CCR1	0.0031	0.0331	UP
CCL5-CCR1	0.0031	0.0331	UP
CCL14-CCR1	0.0031	0.0331	UP
CCL16-CCR1	0.0031	0.0331	UP
CCL23-CCR1	0.0031	0.0331	UP
CCL15-CCR1	0.0031	0.0331	UP
FLT3LG-FLT3	0.0023	0.0331	UP
IL19-IL20RA	0.0023	0.0331	DOWN
INHBE-ACVR1 ACVR2A	0.0042	0.0379	UP
INHBE-ACVR2A ACVR1B	0.0042	0.0379	UP
PDGFA-PDGFRB PDGFRA	0.0042	0.0379	UP
TNFSF13-TNFRSF13B	0.0057	0.0482	UP

Table B.8: Cytokine-Cytokine receptor interaction results in CRC study

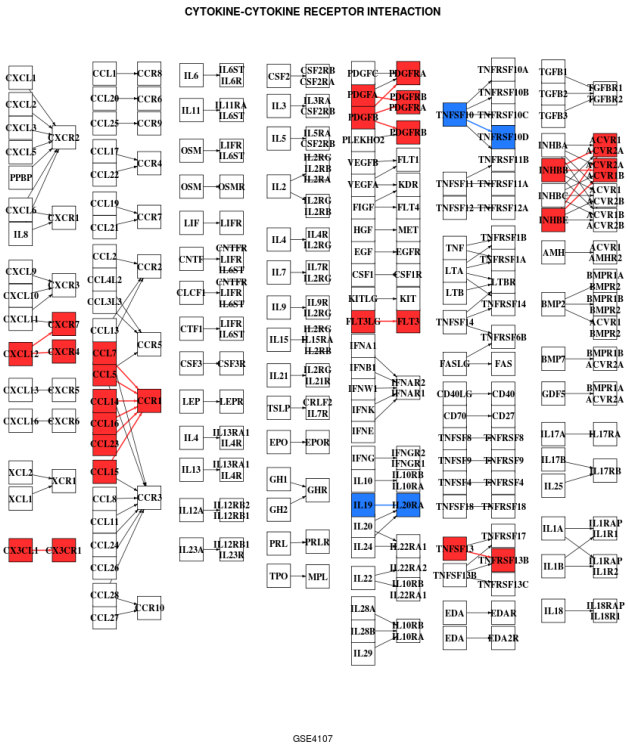


Figure B.7: Cytokine-Cytokine receptor interaction results in CRC study

CHEMOKINE SIGNALING PATHWAY (hsa04062)			
Subpathway	p-value	FDR p.value	UP/DOWN
GRK7-NFKBIA	0.0008	0.0248	UP
CCL26-NFKBIA	0.0012	0.0248	UP
GNB5-NFKBIA	0.0031	0.0451	UP
GNB5-GSK3A	0.0042	0.0456	DOWN
GNB5-FOXO3	0.0057	0.0486	DOWN

Table B.9: Chemokine signaling pathway results in CRC study

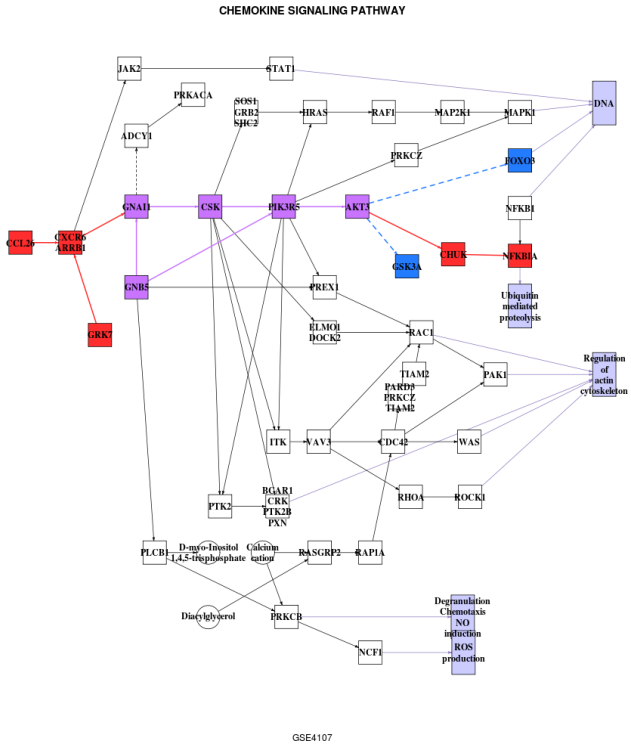


Figure B.8: Chemokine signaling pathway results in CRC study

ECM-RECEPTOR INTERACTION (hsa04512)			
Subpathway	p-value	FDR p.value	UP/DOWN
LAMC3-ITGB1 ITGA7	0.0002	0.0074	UP
FN1-ITGAV ITGB8	0.0002	0.0074	UP
RELN-ITGA3 ITGB1	0.0004	0.0077	UP
FN1-ITGA4 ITGB7	0.0016	0.0267	UP
FN1-ITGB1 ITGA5	0.0042	0.0394	UP
FN1-ITGB1 ITGAV	0.0042	0.0394	UP
FN1-G02170	0.0042	0.0394	UP
FN1-ITGA3 ITGB1	0.0057	0.0408	UP
FN1-SDC1	0.0057	0.0408	UP
FN1-ITGA4 ITGB1	0.0074	0.0484	UP

Table B.10: ECM-receptor interaction results in CRC study

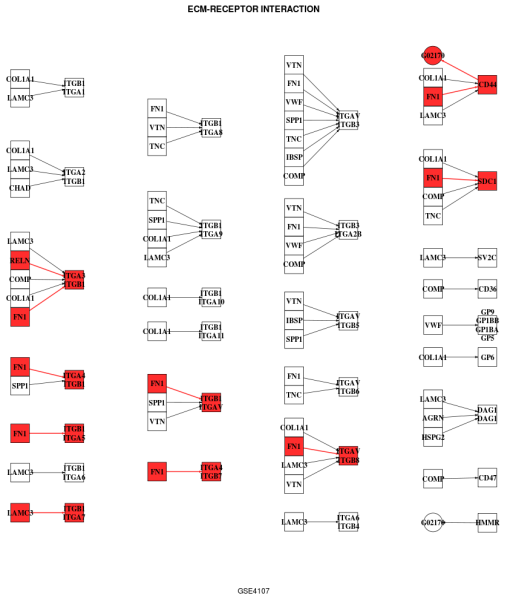


Figure B.9: ECM-receptor interaction results in CRC study

CELL ADHESION MOLECULES (hsa04514)			
Subpathway	p-value	FDR p.value	UP/DOWN
NCAM1-NCAM1	0.0001	0.0031	UP
CDH2-CDH2	0.0001	0.0031	UP
L1CAM-CNTN2 L1CAM CNTN2	0.0012	0.0166	UP
CDH2-CDH2	0.0016	0.0177	UP
SELPLG-GLG1 SELE	0.0023	0.0197	UP
PVR-CD226	0.0057	0.0405	UP
HLA-A-CD8A	0.0074	0.0457	DOWN
CD40LG-CD40	0.0097	0.0463	DOWN
CDH2-CDH2	0.0097	0.0463	DOWN

Table B.11: Cell adhesion molecules results in CRC study

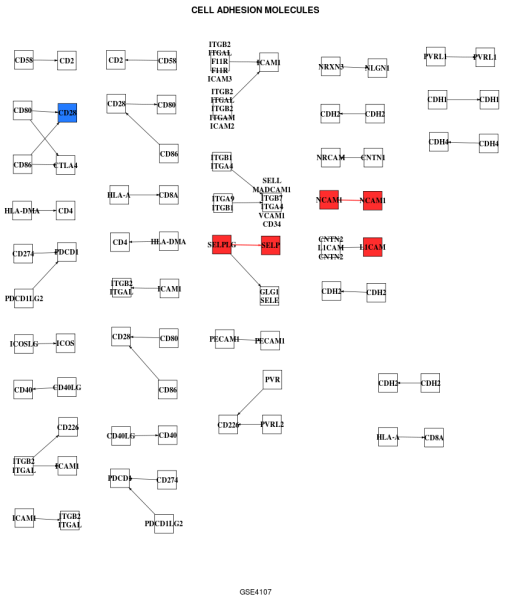


Figure B.10: Cell adhesion molecules results in CRC study

TIGHT JUNCTION (hsa04530)			
Subpathway	p-value	FDR p.value	UP/DOWN
RRAS2-MLLT4	0.0057	0.0311	UP
CSNK2A1-OCLN OCLN	0.0031	0.0311	UP

Table B.12: Tight junction results in CRC study

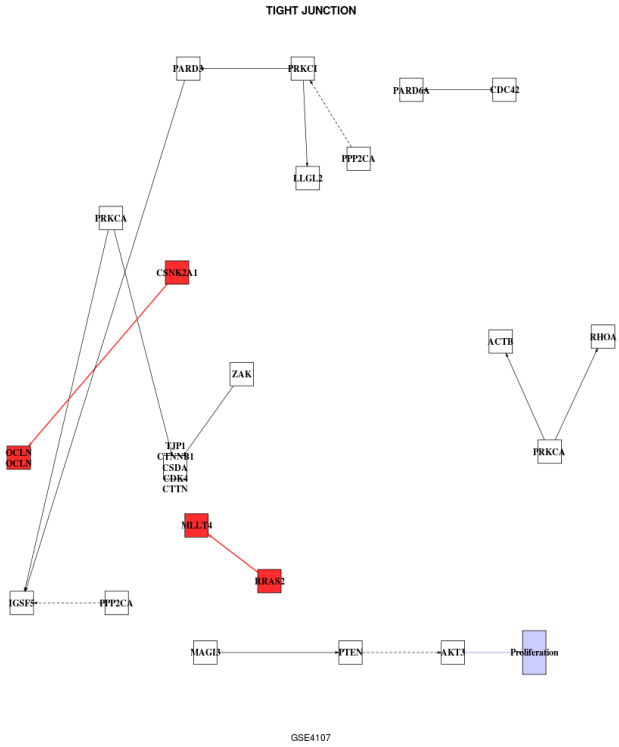


Figure B.11: Tight junction results in CRC study

GAP JUNCTION (hsa04540)			
Subpathway	p-value	FDR p.value	UP/DOWN
DRD1-PRKACA	0.0000	0.0005	UP
ADRB1-PRKACA	0.0001	0.0005	UP
GUCY1A2-PRKG1	0.0057	0.0207	UP
EGF-GJA1	0.0125	0.0343	UP

Table B.13: Gap junction results in CRC study

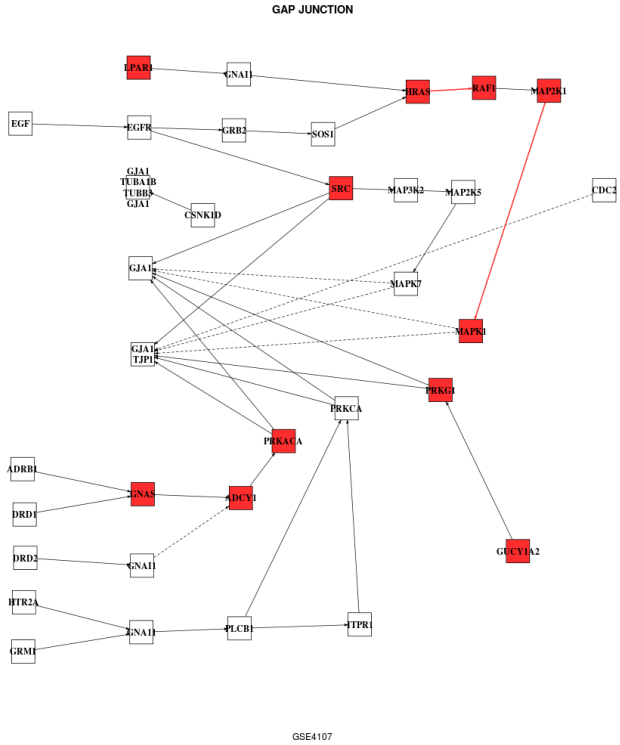


Figure B.12: Gap junction results in CRC study

JAK-STAT SIGNALING PATHWAY (hsa04630)			
Subpathway	p-value	FDR p.value	UP/DOWN
STAM2-MYC	0.0057	0.0130	UP
STAM2-AKT3	0.0074	0.0130	UP
STAM2-PIM1	0.0031	0.0130	UP
STAM2-CCND1	0.0042	0.0130	UP
STAM2-BCL2L1	0.0310	0.0381	UP
STAM2-SPRED1	0.0381	0.0381	UP
STAM2-SPRY3	0.0381	0.0381	UP

Table B.14: JAK-STAT signaling pathway results in colorectal cancer study

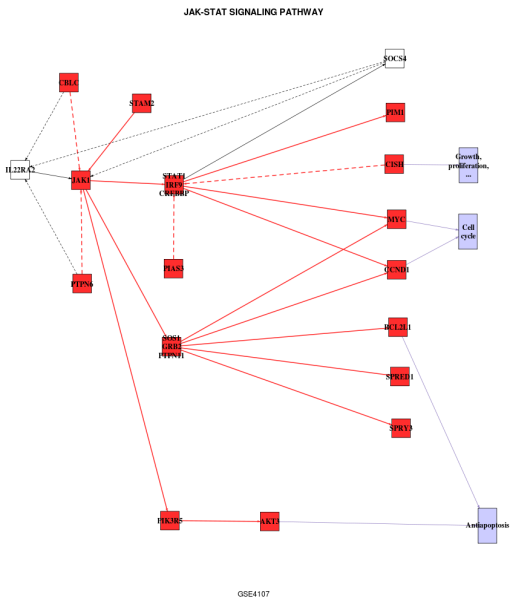


Figure B.13: JAK-STAT signaling pathway results in colorectal cancer study

T CELL RECEPTOR SIGNALING PATHWAY (hsa04660)			
Subpathway	p-value	FDR p.value	UP/DOWN
MAP2K7-FOS	0.0000	0.0003	UP
NCK1 ITK-FOS	0.0001	0.0017	UP
NCK1 ITK-D-myo-Inositol	0.0004	0.0021	UP
NCK1 ITK-CARD11	0.0004	0.0021	UP
NCK1 ITK-IKBKB	0.0005	0.0026	UP
NCK1-PAK4	0.0125	0.0499	DOWN

Table B.15: T cell receptor signaling pathway results in CRC study

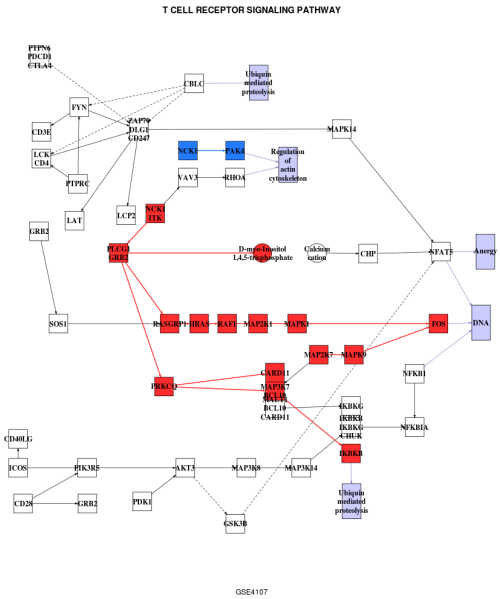


Figure B.14: T cell receptor signaling pathway results in CRC study

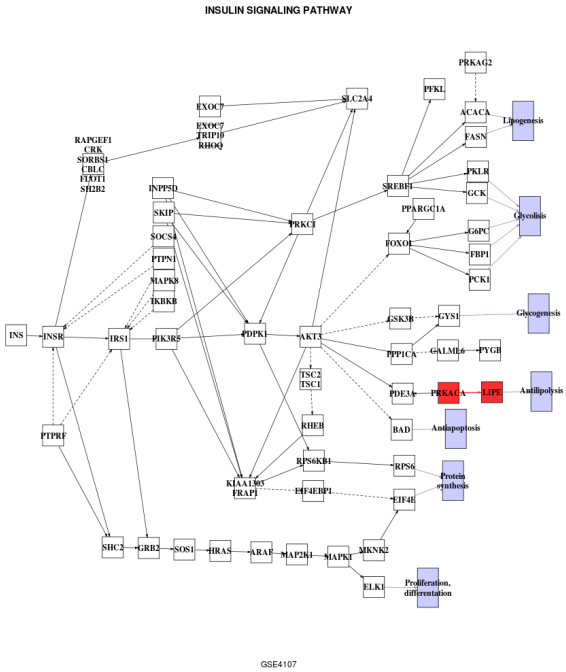


Figure B.15: Insulin signaling pathway results in CRC study

INSULIN SIGNALING PATHWAY (hsa04910)			
Subpathway	p-value	FDR p.value	UP/DOWN
PRKACA-LIPE	4e-04	0.0193	UP

Table B.16: Insulin signaling pathway results in CRC study

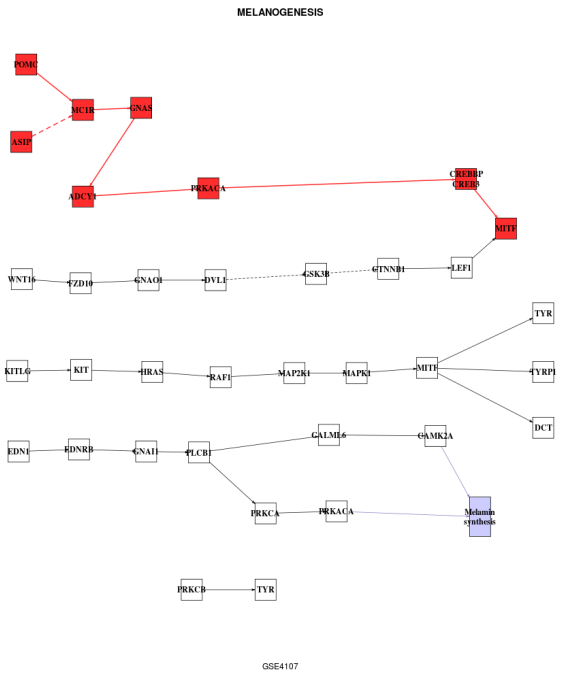


Figure B.16: Melanogenesis in CRC study

MELANOGENESIS (hsa04916)			
Subpathway	p-value	FDR p.value	UP/DOWN
POMC-MITF	0.0057	0.0452	UP

Table B.17: Melanogenesis in CRC study

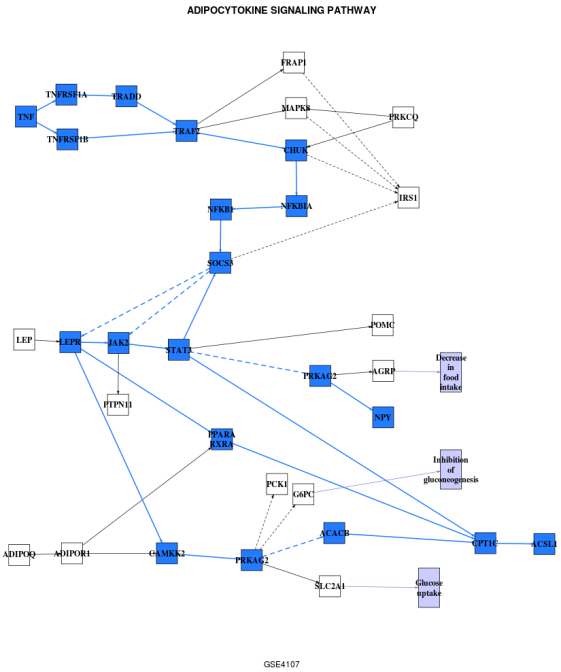


Figure B.17: Adipocytokine signaling pathway results in CRC study

ADIPOCYTOKINE SIGNALING PATHWAY (hsa04920)			
Subpathway	p-value	FDR p.value	UP/DOWN
TNF-ACSL1	0.0000	0.0004	DOWN
TNF-NPY	0.0012	0.0179	DOWN

Table B.18: Adipocytokine signaling pathway results in CRC study

Pathway	alphaMean	alphaVar
Adherens junction	0	0
Measles	0	0
Toll-like receptor signaling pathway	0	0
Vascular smooth muscle contraction	0	0
Viral myocarditis	0	
Acute myeloid leukemia	0	0.01
Glutamatergic synapse	0.01	0
Pathogenic Escherichia coli infection	0	0.01
Prostate cancer	0	0.01
Circadian entrainment	0	0.02
Dopaminergic synapse	0.01	0.01
Epstein-Barr virus infection	0	0.02
Salivary secretion	0.01	0.01
Pancreatic secretion	0	0.03
Cell adhesion molecules (CAMs)	0.05	0
Chronic myeloid leukemia	0	0.05
Gap junction	0	0.05
Hepatitis B	0.01	0.04
Shigellosis	0.05	0
Jak-STAT signaling pathway	0	0.06
Hepatitis C	0.01	0.06
Fructose and mannose metabolism	0.02	0.05
Huntington's disease	0.02	0.06
Serotonergic synapse	0.04	0.04
Cholinergic synapse	0	0.09
NF-kappa B signaling pathway	0.01	0.08
Salmonella infection	0.07	0.02
Cell cycle	0	0.1
GABAergic synapse	0.05	0.05
Natural killer cell mediated cytotoxicity	0	0.1
Neuroactive ligand-receptor interaction	0.08	0.02
Nicotinate and nicotinamide metabolism	0.02	0.09
Oocyte meiosis	0.05	0.06
Antigen processing and presentation	0	0.12
Dilated cardiomyopathy	0.01	0.11
Glycerophospholipid metabolism	0	0.12

HIF-1 signaling pathway	0	0.12
Leukocyte transendothelial migration	0	0.12
Colorectal cancer	0.01	0.12
Influenza A	0	0.14
Retrograde endocannabinoid signaling	0.03	0.11
Leishmaniasis	0	0.16
Apoptosis	0	0.17
Herpes simplex infection	0	0.18
Gastric acid secretion	0.01	0.18
RIG-I-like receptor signaling pathway	0	0.19
African trypanosomiasis	0.01	0.2
Cocaine addiction	0.04	0.17
Adipocytokine signaling pathway	0	0.22
Long-term depression	0.03	0.19
Osteoclast differentiation	0.01	0.21
Tight junction	0	0.22
Amyotrophic lateral sclerosis (ALS)	0.05	0.17
Progesterone-mediated oocyte maturation	0.02	0.21
mTOR signaling pathway	0	0.23
Prion diseases	0.01	0.22
Bacterial invasion of epithelial cells	0.01	0.23
Carbohydrate digestion and absorption	0	0.25
Pentose phosphate pathway	0.25	0.03
Glioma	0	0.29
Melanoma	0.02	0.27
Small cell lung cancer	0.02	0.27
B cell receptor signaling pathway	0.03	0.27
Inositol phosphate metabolism	0.04	0.27
ECM-receptor interaction	0	0.33
Bladder cancer	0.04	0.3
Alzheimer's disease	0	0.36
Amino sugar and nucleotide sugar metabolism	0	0.38
Morphine addiction	0.01	0.41
Fc gamma R-mediated phagocytosis	0	0.43
Axon guidance	0	0.46
Insulin signaling pathway	0.01	0.49
Pertussis	0	0.52
Thyroid cancer	0.02	0.52
Wnt signaling pathway	0	0.55

Fc epsilon RI signaling pathway	0.02	0.57
Glycosphingolipid biosynthesis - lacto and neolacto series	0.62	0
Neurotrophin signaling pathway	0	0.62
TGF-beta signaling pathway	0	0.63
Toxoplasmosis	0	0.66
Endometrial cancer	0	0.68
NOD-like receptor signaling pathway	0	0.71
Endocrine and other factor-regulated calcium reabsorption	0.03	0.71
VEGF signaling pathway	0.05	0.72
ErbB signaling pathway	0	0.92

Table B.19: Significant results which came from GraphiteWeb in the CRC dataset

Chapter C

Abbreviations

AML Acute Myeloid Leukemia

ASW African Ancestry in Southwest US

ATM adipose tissue macrophages

AUC area under the curve

BIC Bayesian Information Criterion

CAMDA Critical Assessment of Massive Data Analysis

CCLE Cancer Cell Line Encyclopedia

CEU Utah residents with Northern and Western-European ancestry

CFS correlation-based feature selection

CHB Han Chinese in Beijing

CHS Southern Han Chinese

CLM Colombian in Medellin

CPG Cancer Genome Project

CRAN Comprehensive R Archive Network

CRC colorectal cancer

DNA Deoxyribonucleic acid

DREAM The Dialogue for Reverse Engineering Assessments and Methods

FA Fanconi Anemia

FCS Functional Class Scoring

FDR False Discovery Rate

FIN Finnish in Finland

GBR British in England and Scotland

GEO Gene Expression Omnibus Database

HSC hematopoietic stem cells

IBS Iberian populations in Spain

IC50 half-maximal inhibitory concentration

IF impact factor

JPT Japanese in Tokyo

KEGG Kyoto Encyclopedia of Genes and Genomes

KNN K-Nearest neighbor

KGML KEGG Markup Language

LWK Luhya in Webuye

MCC Mathews correlation coefficient

miRNA microRNA

MM Mismatch probe

mRNA Messenger RNA

MRP masters degree research project

MXL Mexican Ancestry in Los Angeles

NCBI National Center for Biotechnology Information

ORA Over-Representation Analysis

PANP presence-absence calls with negative probesets

PBMC peripheral blood mononuclear cells

PCC proportion of correct classification

PCR Polymerase Chain Reaction

PM Perfect match probe

PolyPhen Polymorphism Phenotyping

PPI Protein-protein interaction

PUR Puerto Rican in Puerto Rico

RMSE root mean square error

RNA Ribonucleic acid

RNA-seq RNA Sequencing

RP research project

RT-PCR Reverse Trancription PCR

SIF Sorting Intolerant From Tolerant

SNV single nucleotide variation

SVM Support vector machine

TSI Toscani in Italy

TF Transcription Factor

XML Extensible Markup Language

YRI Yoruba in Ibadan

Aminoacids abbreviations:

Ala Alanine

Gin Glutamine

Leu Leucine

Ser Serine

Arg Arginine

Glu Glutamic acid

Lys Lysine

Thr Threonine

Asn Asparagine

Gly Glycine

Met Methionine

Trp Trytophane

Asp Aspartic acid

His Histidine

Phe Phenylalanine

Tyr Tyrosine

Cys Cysteine

Ile Isoleucine

Pro Proline

Val Valine